A new adaptive multiscale finite element method with applications to high contrast interface problems

Millward, Raymond

Award date:
2011

Awarding institution:
University of Bath

A new adaptive multiscale finite element method with applications to high contrast interface problems

submitted by

Raymond R. Millward

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

May 2011

Summary

In this thesis we show that the finite element error for the high contrast elliptic interface problem is independent of the contrast in the material coefficient under certain assumptions. The error estimate is proved using a particularly technical proof with construction of a specific function from the finite dimensional space of piecewise linear functions.

We review the multiscale finite element method of Chu, Graham and Hou to give clearer insight. We present some generalisations to extend their work on a priori contrast independent local boundary conditions, which are then used to find multiscale basis functions by solving a set of local problems. We make use of their regularity result to prove a new relative error estimate for both the standard finte element method and the multiscale finite element method that is completely coefficient independent.

The analytical results we explore in this thesis require a complicated construction. To avoid this we present an adaptive multiscale finite element method as an enhancement to the adaptive local-global method of Durlofsky, Efendiev and Ginting. We show numerically that this adaptive method converges optimally as if the coefficient were smooth even in the presence of singularities as well as in the case of a realisation of a random field.

The novel application of this thesis is where the adaptive multiscale finite element method has been applied to the linear elasticity problem arising from the structural optimisation process in mechanical engineering. We show that a much smoother sensitivity profile is achieved along the edges of a structure with the adaptive method and no additional heuristic smoothing techniques are needed.

We finally show that the new adaptive method can be efficiently implemented in parallel and the processing time scales well as the number of processors increases. The biggest advantage of the multiscale method is that the basis functions can be repeatedly used for additional problems with the same high contrast material coefficient.

Acknowledgements

First and foremost, I would like to thank my PhD supervisor Ivan Graham whose direction, patience and supportive nature has taught me a great deal about the hard work needed for research and the value of being a well-rounded mathematician. Without his guidance from beginning to end this thesis would not have been possible, thank you very much Ivan.

Secondly I want to thank my wife Sam for being there with me through the highs and the lows, for keeping me motivated and for pushing me to acheive the best that I can. In particular thank you for your support over the last few months with the long working weeks, I am forever indebted to you. I would also like to thank all of my family, particularly my dad, for all of their help, support and encouragement throughout my years in Bath. I would like to dedicate this thesis to the memory of my grandparents, who helped me to grow but also provided home comforts during my undergraduate degree, and also to the memory of my mum.

Furthermore, I would like to thank all the friends and people that have influenced me over my time in Bath. To my housemates Chris, Esther, Jeff thank you for all the great times and fun together. To my officemates and other PhD students for our helpful discussions and office banter; Fynn, Tania, Giampiero, Natalya, Jane, Adam, Dave and Phil. To all the academic staff who have asked questions and helped to improve my work over the past three and a half years, in particular Rob and Adrian for their feedback in the progress meetings and in seminars.

I would also like to acknowledge the helpful discussions with Peter Dunning and Dr Alicia Kim regarding the structural optimisation process, especially to Peter and Phil for some of the optimisation images in this thesis.

Special thanks go to all the administration staff and computer support staff in the Department for all their help over the years whether seen or unseen. Also thank you to the EPSRC for the funding of this project.

Contents

1 Introduction

1.1 The subject of the thesis

In nature complex systems operate on many scales in space and time. As scientific models become more complex it is apparent that these different length scales must be included to capture the true behaviour of a system accurately. Multiscale modelling seeks to introduce methods that can capture, utilise and link scales together but with an amount of work that remains constant as the smallest scale decreases. An example of multiscale modelling comes from physics and the determination of material properties.

Multiscale modelling is a vast field of research with significant study over the last ten years. What we examine in this thesis is part of this field covering multiscale finite element methods (FEMs) for elliptic PDEs. The fine scales make standard FEMs converge poorly with respect to the size of the elements used due to a loss in regularity. This poor convergence is worse if the fine scale properties vary significantly in size, e.g. if a thermal insulator (with high thermal resistance) like ceramic, is next to a thermal conductor (with low thermal resistance) like metal.

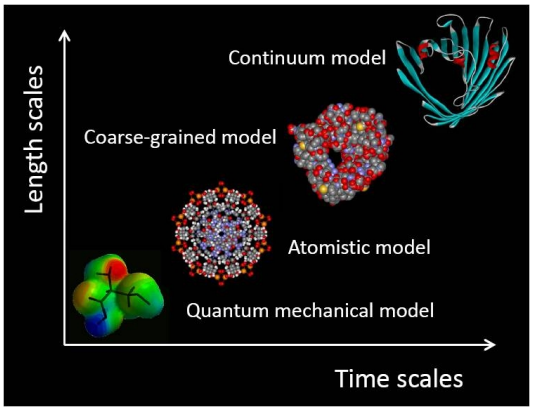


Figure 1-1: *The scales of multiscale modelling. [Courtesy of A. Heyden [46]]*

This thesis is concerned with approximating the solution, u, to a second order elliptic PDE in the weak form

$$a_\Omega(u, v) = L_\Omega(v) \tag{1.1}$$

where a_Ω is a bilinear form that depends on a coefficient $\mathcal{A}(x)$ and domain Ω, see

Section 2.1. $\mathcal{A}(x)$ is known as the permeability function from fluid flow through porous media. $\mathcal{A}(x)$ can be almost anything, provided that the bilinear form remains bounded and coercive; it could range from being smooth to being heterogeneous. In this thesis we are interested in the subset of elliptic PDEs where the ratio between maximum and minimum values of $\mathcal{A}(x)$ is very large, five to ten orders of magnitude. This ratio is known as the "contrast" and there are few results about how the FE error depends on $\mathcal{A}(x)$ and the contrast.

What the FEM seeks to do is approximate the solution of (1.1) on a mesh $\mathcal{T}_H(\Omega)$, where a mesh is a set of simplices that cover Ω and the maximum element diameter is H. This is where multiscale modelling comes in. As the features of the permeability field $\mathcal{A}(x)$ shrink (e.g. they are of order h in size) it is important to model these on a coarse scale (e.g. on a mesh $\mathcal{T}_H(\Omega)$ where $h << H$) but retain the accuracy of modelling all the smaller components.

This approximation uses a finite dimensional subspace, V_H, of the solution space (e.g. if $u \in H^1(\Omega)$ then take a $V_H \subset H^1(\Omega)$) and solving the FE problem

$$a_\Omega(u_H, v_H) = L_\Omega(v_H) \tag{1.2}$$

for any $v_H \in V_H$ to get the FE approximation u_H. The quantity of interest is the FE error $u - u_H$ measured in various norms, most notably the energy norm

$$|u - u_H|_{H^1(\Omega),\mathcal{A}} = a_\Omega(u - u_H, u - u_H)^{\frac{1}{2}}. \tag{1.3}$$

The art of FEMs is the choice of space V_H, the many methods arise from choosing a different V_H and a set of basis functions that span V_H. The choice of V_H may lead to a smaller and thus better FE error. A standard FEM uses the space of continuous functions that are polynomial (e.g. linear) on the simplices of the mesh $\mathcal{T}_H(\Omega)$.

Producing a priori error bounds for all elliptic PDEs with heterogeneous coefficients is a difficult problem. To make proving error bounds more tractable we consider a subset of these problems known as interface problems. We consider a domain Ω that contains a finite number of inclusions Ω_i, $i = 1, ..., m$. We restrict the permeability field to smooth slowly varying functions in each inclusion but that can jump across the interface between inclusions. For example the shades of grey in the radioactive waste vault example (Figure 1-2(a)) define inclusions.

The restriction to interface problems is only required to prove theoretical results in Chapters 2 and 3. We show in Chapter 2 that the standard FE error is bounded by

$$|u - u_H|_{H^1(\Omega),\mathcal{A}} \leq CH^{\frac{1}{2}} \tag{1.4}$$

for a constant C when $\mathcal{A}(x)$ is a discontinuous permeability field. Crucially we show in Chapter 2 how C is independent of the contrast in $\mathcal{A}(x)$ and in Chapter 3 how the relative FE error

$$\frac{|u - u_H|_{H^1(\Omega),\mathcal{A}}}{\max_i |u|_{H^2(\Omega_i)}} \leq C' H^{\frac{1}{2}} \tag{1.5}$$

is independent of the coefficient. However, if $\mathcal{A}(x)$ were smooth then we would expect the FE error to have the bound

$$|u - u_H|_{H^1(\Omega),\mathcal{A}} \leq CH \tag{1.6}$$

since u would have sufficient regularity to be in $H^2(\Omega)$. This drop in convergence rate makes the standard FEM a poor choice for elliptic PDEs with discontinuous coefficients.

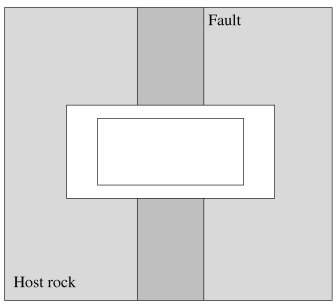
Ideally we want to have a FEM that gives a finite element error of order H like (1.6) even in the presence of discontinuous coefficients. This is where the idea of multiscale finite elements comes in. There are many methods that try to incorporate the fine scales into V_H. Methods like the extended FEM (XFEM) by Moes, Dolbow and Belytschko [69] and the residual free bubble method (RFBM) by Brezzi et al [22, 21] seek to enrich V_H with additional basis functions that better match the shape of the solution. Other methods upscale $\mathcal{A}(x)$ replacing it by a constant on each simplex of the mesh $\mathcal{T}_H(\Omega)$ and solving the FE problem (1.2) with this upscaled field. The fine scale information incorporated into coarse scale enrichment functions or upscaled permeabilities then interact through the variational form (1.2).

After Chapter 2 this thesis will focus on another class of multiscale methods that use multiscale basis functions. The idea is to incorporate the fine scale information into the basis functions themselves by solving a local problem, based on a homogeneous version of (1.1), around each simplex of $\mathcal{T}_H(\Omega)$.

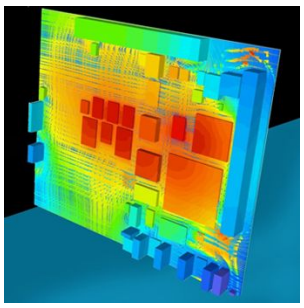
Examples of high contrast interface problems arise in many areas of engineering. Most significantly in modelling groundwater flow. This is increasingly important with the resurgence of nuclear power and new nuclear waste storage facilities (Figure 1-2(a)). As a consequence, it is important to know how environmentally secure these facilities are through modelling [90]. Additional questions arise when a geological fault is allowed to run through the structure. All of these questions require an accurate solution of the interface problem to ensure a confident analysis and informed decision making.

On an engineering level it is increasingly important to model smaller features when doing heat transfer analysis of circuit boards in electronics [44]. As devices shrink it

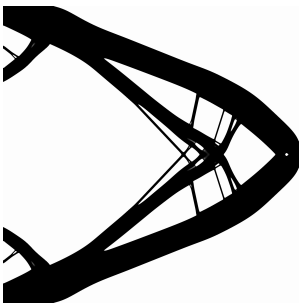
becomes important to take into account the differences in materials present across the design (Figure 1-2(b)).



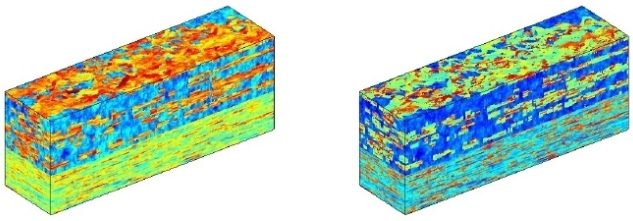
(a) Groundwater flow.



(b) Heat transfer in electronics. (Courtesy of Mentor Graphics Mechanical Analysis Division [42])



(c) Structural optimisation.



(d) SPE10 upscaling benchmark. (Courtesy of K.-A. Lie [63])

Figure 1-2: *Examples of high contrast elliptic PDEs with discontinuous coefficients.*

Although the restriction to interface problems for theoretical error estimates seems limiting in Chapters 2 and 3 it actually leads to a novel application of multiscale FEMs to the field of structural optimisation that we will cover in more detail in Chapter 5. As computer modelling of aircraft, motorsport and building structures becomes widespread it is increasingly important to have very detailed accurate solutions as the strength of a material comes from its microstructure. This gives us a multiscale problem as the scale of the design, e.g. a building, is vastly bigger than the scale of the microstructure.

A mechanical engineer seeks to analyse the stress and strain of a stucture being designed. They interpret this to find when and how a structure is likely to fail. Typically a structure will fail where there are concentrations of stress or strain, so an even distribution of stress and strain is desired. If there are problems with a structure it may get re-designed to make it stronger.

Structural optimisation seeks to automate this process of redesign to find the best structure under loading conditions and constraints, e.g. the best cantilever to support a load hanging from one end and fixed to a wall at the other (Figure. 1-2(c)). In certain

special cases this optimal structure can be determined by solving one linear elasticity problem, typically though it is done iteratively.

This presents a significant modelling challenge. A lot of microstructure provides a strong macrostructure but with a reduced amount of material. However the mesh $\mathcal{T}_H(\Omega_0)$ needs to be very fine to resolve all of these fine scales. Therefore a complex shape must be re-meshed several hundred times. Instead recent work has considered fixed mesh approaches. A larger design space is considered, Ω, in which a binary material coefficient $\mathcal{A}(x)$ is considered corresponding to material and no material. A fixed mesh $\mathcal{T}_H(\Omega)$ is set over the domain and then it is $\mathcal{A}(x)$ that alters between iterations. The standard FEM converges poorly if the boundaries of the structure run through the inside of elements. Multiscale methods offer a way of improving the convergence rate as well as avoiding the complex re-meshing.

The final example returns back to (1.1) to look at problems with heterogeneous permeability fields in rock structures. The problem comes from the Society of Petroleum Engineers and Figure 1-2(d) shows the permeability field for SPE10, a benchmark to test upgridding and upscaling examples against. We give a model problem motivated by fluid flow in a porous medium. It makes several assumptions; the flow is incompressible, the fluid has constant density, overburden and atmospheric pressure are constant, the groundwater has a Reynolds number less than one (slowly flowing). These restrictions still give a good representation of pressure heads in an aquifer. We also restrict to a steady-state system, thus the general diffusion equation from porous media flow simplifies to a second order elliptic PDE as shown next. For an explanation of the physics of fluid flow in porous media see [84] by Wang and Anderson.

Example 1.1. *The flow problem is derived by considering Darcy's law for fluid flow through a porous media along with conservation of mass in a fixed volume. We consider the steady state problem which gives us the simplified conservation of mass law*

$$\nabla \cdot q - f = 0 \tag{1.7}$$

where q is the Darcy flux, the fluid discharge per unit area, and f is the source/sink term for fluid generation/loss. The Darcy flux is then given by Darcy's law as

$$q = -\mathcal{A}(x)\nabla u \tag{1.8}$$

where $\mathcal{A}(x)$ is the permeability field and ∇u is the pressure gradient. These combine to give the classical single phase flow equation

$$\nabla \cdot (-\mathcal{A}(x)\nabla u) - f = 0 \tag{1.9}$$

more readily written in the form below along with boundary conditions

$$\begin{aligned} -\nabla \cdot (\mathcal{A}(x)\nabla u) &= f \qquad \text{on } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{1.10}$$

For a discontinuous field $\mathcal{A}(x)$ this problem particularly must be solved in the weak form corresponding to Problem 2.2 given in Chapter 2 because the term $\nabla \cdot (-\mathcal{A}(x)\nabla u)$ does not exist at the points where $\mathcal{A}(x)$ jumps. Note that BVPs also require a weak formulation to be solved in order to incorporate discontinuous boundary conditions, Neumann boundary conditions and discontinuous load functions.

In this thesis we consider (1.1) applied to such high contrast problems and consider how the contrast affects the convergence of the standard FEM, giving theoretical results for a subset of interface problems. Then we will look at convergence results for a multiscale FEM devised in [27] and extend these. Following that we will consider an adaptive multiscale method for (1.1), an extension of the method by Durlofsky, Efendiev and Ginting in [36], for defining a set of multiscale basis functions. Finally we will apply this new adaptive multiscale FEM to structural optimisation and describe a parallel version of the algorithm. Before that we take a look at previous work in each of these areas to see how our work fits into the wider context.

1.2 Literature review

1.2.1 Early multiscale methods, convergence and contrast dependence

The idea of the multiscale finite element method, whereby better basis functions are found by solving a local homogeneous PDE with specific boundary conditions has a large literature.

1.2.1.1 Optimal order convergence

Work that aims at showing optimal order convergence for 2-dimensional interface problems can be found in [25] by Chen and Zou. Here they approximated a smooth C^2 interface Γ by a polygon Γ_H with nodes on Γ and sides of at most H in length. They then create a mesh $\mathcal{T}_H$ where the elements have at most two nodes on Γ_H and then solve the finite element problem. This method of resolving the interface means Chen and Zou can use the fact that for an element $\tau \in \mathcal{T}_H$ that intersects Γ will have an intersection area

$$\text{meas}(\tau \cap \Omega_1) \leq CH_\tau^3 \qquad \text{or} \qquad \text{meas}(\tau \cap \Omega_2) \leq CH_\tau^3$$

where Ω_1 and Ω_2 are the two inclusions that Γ is an interface for. While this produces optimal order finite element error estimates it places complicated restrictions on how the mesh is set up, also the error estimates in [25] do not explicitly state the dependence on the coefficient $\mathcal{A}(x)$, this dependence is simply absorbed into the constant of proportionality.

The closest work to that presented in this thesis in Chapter 2 is that of Li, Melenk, Wohlmuth and Zou [59]. Here they presented approximation error bounds for the standard FEM applied to the two- and three-dimensional elliptic interface problem

$$\int_\Omega \mathcal{A}(x)\nabla u \cdot \nabla \phi \;\; dx = \int_\Omega f\phi \;\; dx \qquad \text{for all } \phi \in H_0^1(\Omega) \tag{1.11}$$

with $u = 0$ on $\partial\Omega$. They present hp-finite element error estimates that combine error estimates based on the size and shape of elements in the mesh $\mathcal{T}_H(\Omega)$ (the h-finite element error based on the maximum element size H) and the order of polynomial used in the finite dimensional spaces

$$V_H^{\mathbb{P}_p} = \left\{ v \in C^0(\Omega) \mid \quad v|_\tau \in \mathbb{P}_p \text{ for all } \tau \in \mathcal{T}_H(\Omega) \right\}$$

where $\mathbb{P}_p$ is the space of polynomials up to order p (the p-finite element error based on the maximum polynomial order p). They showed that optimal order convergence could be obtained (estimates of order H^p, like (1.6) where $p = 1$) provided the finite element mesh sufficiently resolved the interface. To explain this we consider an interface Γ dividing a domain Ω into Ω_1 and Ω_2. Then for any element τ of the mesh $\mathcal{T}_H(\Omega)$ that cuts the interface, $\text{int}(\tau) \cap \Gamma \neq \emptyset$, define its minimum intersection distance into an inclusion by

$$\delta_\tau = \min_{i=1,2} \left\{ \max \left\{ \text{dist}(x, \Gamma \cap \tau) \mid x \in \tau \cap \Omega_i. \right\} \right\}$$

So an element is mostly in one inclusion but the part in the other inclusion is only of size δ_τ. Then define

$$\delta = \max_{\tau \in \mathcal{T}_H(\Omega)} \delta_\tau,$$

which then leads to the definition of the mesh sufficiently resolving the interface. It is sufficiently resolved if δ is of order H^{2p} for mesh size H and approximating polynomial order p. This is important because it shows that the mesh does not have to resolve

the interface exactly ($\delta = 0$) which is impossible for a curved interface and standard triangular elements. We reinforce their error estimates in Chapter 2 of this thesis using piecewise linear continuous functions $V_H^{\mathbb{P}_1}$ for the standard FEM, however crucially we extend it to the case of high contrast coefficients. Rather than looking at rates with respect to the mesh size H we are looking at the dependence of error estimates on the coefficient $\mathcal{A}$.

Plum and Wieners have managed to show optimal a priori convergence rates in arbitrary dimensions but under certain very specific conditions. The major condition being the existence of an interpolation operator Π_H into the finite element space $V_H \subset H_0^1(\Omega)$ satisfying

$$\|\nabla (v - \Pi_H(v))\|_{L_2(\Omega_k)} \leq CH \left\|D^2 v\right\|_{L_2(\Omega_k)} \qquad \text{for all } v \in H^2(\Omega_k)$$

where Ω is the union of the non-overlapping subdomains Ω_k, $k = 1, ..., m$. This is very restrictive though as for standard hp-finite elements this only happens when the mesh resolves the interface, i.e. $\tau \subset \Omega_k$ for any element τ of the finite element mesh $\mathcal{T}_H(\Omega)$.

1.2.1.2 Contrast independence

The crucial point, specifically concerning the purpose of this thesis, is that the error bounds in all of the above works have a constant that is dependent on the coefficient $\mathcal{A}(x)$ but more importantly also the contrast of $\mathcal{A}(x)$ where the contrast is defined by

$$\frac{\max \mathcal{A}(x)}{\min \mathcal{A}(x)}. \tag{1.12}$$

Only [17] by Babuška and Osborn has a finite element error bound associated with the solution built from the harmonic average of $\mathcal{A}(x)$ (see (1.14)) that is independent of the maximum of $\mathcal{A}(x)$, they do not consider a relative estimate of the form (1.5) to show that it is also independent of the minimum. However, showing independence from the contrast in the coefficient in the constant of (1.6) was not the aim of their work, Babuška et al were simply trying to show how the rate of convergence with respect to the mesh size H could be improved from the standard FEM results. Significantly the coefficient independent result in [17] was only shown for the 1D interface problem (1.13); proving coefficient independence is much harder in 2D.

While [18] by Bernardi and Verfürth is mostly concerned with a posteriori error estimates it does contain a section on a priori estimates for the 2D interface problem (1.11). Bernardi and Verfürth showed that you could get optimal convergence independent of

the contrast in Section 2.c of [18], however it still depends on the coefficient $\mathcal{A}(x)$ itself. So when $\mathcal{A}(x)$ is very large in an inclusion the error bound becomes very poor. It is also unclear how the H^{1+s} norm of the gradient of the solution ∇u, $0 \leq s \leq 1$, in the right hand side of their error estimates (Theorem 2.5 in [18]) depends on the coefficient $\mathcal{A}(x)$. Finally and most importantly they made the restrictive assumption that the interface was resolved by the mesh.

1.2.1.3 Regularity results for contrast independence

What we will show in Chapter 2 is that the standard FEM error in approximation is in fact independent of the maximum of $\mathcal{A}(x)$ and then using a bound on the relative error (the left hand side of (1.5)) we will show that the error in approximation is independent of the contrast in $\mathcal{A}(x)$. The key to being able to achieve this extension to the current results comes from a novel regularity result in the appendix of [27] that gives bounds on the seminorms of the solution that are explicit in the coefficient $\mathcal{A}(x)$. While this is only done for a single inclusion we will extend it to the case of multiple inclusions in Chapter 3 thus allowing a relative error estimate to be constructed. Babuska, Caloz and Osborn introduced a regularity result in [16] but it is unclear exactly how the constants depend on the contrast and it relies on the coefficient being unidirectional (Figure 1-3). The earlier work of Huang and Zou in [51] gives a partial result in the same direction as [27]. Consider a domain $\Omega = \Omega_1 \cup \Omega_2$ where Ω_1 is an inclusion inside Ω that does not touch the boundary $\partial\Omega$, also suppose the coefficient $\mathcal{A}(x)$ is piecewise constant such that $\mathcal{A}|_{\Omega_1} := \mathcal{A}_1 > \mathcal{A}_2 =: \mathcal{A}|_{\Omega_2}$. Huang and Zou showed coefficient explicit bounds on the full H^2 norms of u for the surrounding material Ω_2,

$$\|u\|_{H^2(\Omega_2)} \lesssim \frac{1}{\mathcal{A}_2} \|f\|_{L_2(\Omega)},$$

but is not explicit for the island inclusion Ω_1 inside the domain giving only

$$\|u\|_{H^2(\Omega_1)} \lesssim \|f\|_{L_2(\Omega)}.$$

The coefficient explicit seminorm bounds from [27] are essential for proving coefficient independent finite element error estimates.

1.2.1.4 Historical context

The early form of this method started with Babuška [11] where the error in approximation for the standard finite element method (FEM) was shown to be very poor for

the one-dimensional elliptic interface problem

$$\int_a^b \mathcal{A}(x)\frac{\partial u}{\partial x}\frac{\partial \phi}{\partial x}\ dx = \int_a^b f\phi\ dx \qquad \text{for all } \phi \in H_0^1([a,b]) \tag{1.13}$$

when the discontinuity in $\mathcal{A}(x)$ was inside an element. This means the jump in $\mathcal{A}(x)$ was in $(a_i,\ a_{i+1})$ for a partition $\mathcal{T}_H([a,b]) = \{[a_j,\ a_{j+1}]\,|\,a_j < a_{j+1} \quad \text{for all } j = 0, ..., N\}$. Babuška continued to investigate the rapidly jumping coefficient within a homogenisation setting in [12, 13, 14].

In 1983 Babuška and Osborn [17] introduced the idea of the generalised finite element method (GFEM) where the standard method is a special case, the generalised method being a combination of the standard FEM and the partition of unity method (where a set of functions that span the original space are defined with only finitely many being non-zero at each point and all the functions summing to 1 at each point). They considered only one-dimensional interface problems like (1.13) but with rough coefficients (meaning that no matter how fine the mesh $\mathcal{T}_H([a,b])$ got, with arbitrarily large N, the coefficient function $\mathcal{A}(x)$ always had a discontinuity inside at least one element). They showed that this problem would not converge for the standard FEM, i.e. $u - u_H \nrightarrow 0$ as $H \to 0$, but instead by solving the problem with what is known as the 'harmonic average' of $\mathcal{A}(x)$,

$$\mathcal{A}_{\text{harmonic}}|_{[a_j,a_{j+1}]} = \left(\frac{\int_{a_j}^{a_{j+1}} \frac{1}{\mathcal{A}(x)}dx}{a_{j+1}-a_j}\right)^{-1}, \tag{1.14}$$

instead of $\mathcal{A}(x)$ they could obtain a good approximation that converged very well. These results were then extended by Babuška, Caloz and Osborn [16] to two-dimensional second order elliptic interface problems but restricted to the case that the coefficient $\mathcal{A}(x)$ is uni-directional, e.g. $\mathcal{A}(x_1, x_2) = \mathcal{A}_1(x_1)$. This idea also applies to curvilinear coordinates for example if the coefficient $\mathcal{A}$ only depends on the radius as in Figure 1-3.

The method does rely on being able to map a curvilinear triangle back to the reference triangle thus transforming the special basis functions that utilize the 'harmonic average' (1.14) into polynomials. This also transforms the unknown function into a smooth function thus allowing the theory with smooth coefficients to be used. The benefit is that they do obtain optimal error bounds with respect to the mesh size H, in the sense that they get estimates like (1.6) as if there were no loss of regularity from the discontinuities of $\mathcal{A}(x)$ being inside mesh elements. More recently in 2004 Babuška, Banerjee and Osborn [15] gave a summary of work so far with the Generalised FEM and includes the general two-dimensional second order elliptic interface problem (1.1).

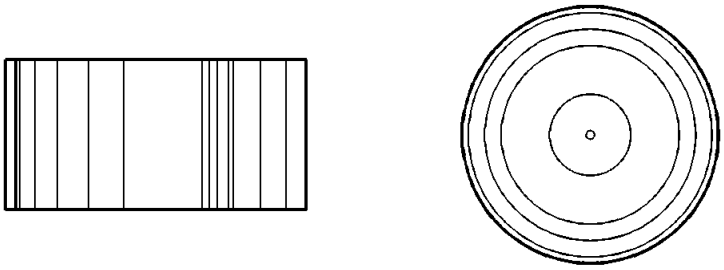


Figure 1-3: *Examples of unidirectional composites in [16].*

They briefly mention the difficulty of selecting local approximation spaces V_H when the solution has a singularity, listing the case when the coefficient $\mathcal{A}(x)$ is piecewise smooth with jumps as we will consider in Chapter 2, but they do not produce any error estimates for this case.

1.2.2 Advances in multiscale methods

1.2.2.1 Historical upscaling techniques

In order to combat the poor convergence shown for the standard finite element method applied to multiscale problems, many techniques have been introduced. In many cases there is far more data about a model than can be incorporated into a finite element discretisation, for example data about the permeability of rock in an oil field. Figure 1-4 shows a typical fine scale distribution of permeability information (left). There has been a lot of work on the idea of 'upscaling' that data to a coarser finite element mesh, which gives an effective permeability field (Figure 1-4 right) that is suitable for computer simulations. The upscaling may be done in many different ways, for example just taking the arithmetic average over a coarse element

$$\mathcal{A}_{\text{upscaled}}|_{\tau} = \frac{\int_{\tau} \mathcal{A}(x)\ dx}{\int_{\tau} dx}$$

or even the harmonic average

$$\mathcal{A}_{\text{upscaled}}|_{\tau} = \left(\frac{\int_{\tau} \frac{1}{\mathcal{A}(x)} dx}{\int_{\tau} dx} \right)^{-1} .$$

This is linked with the ideas of Babuška, and Osborn [17] who used the harmonic average of the coefficient on a coarser mesh in 1D. Other upscaling work can be found

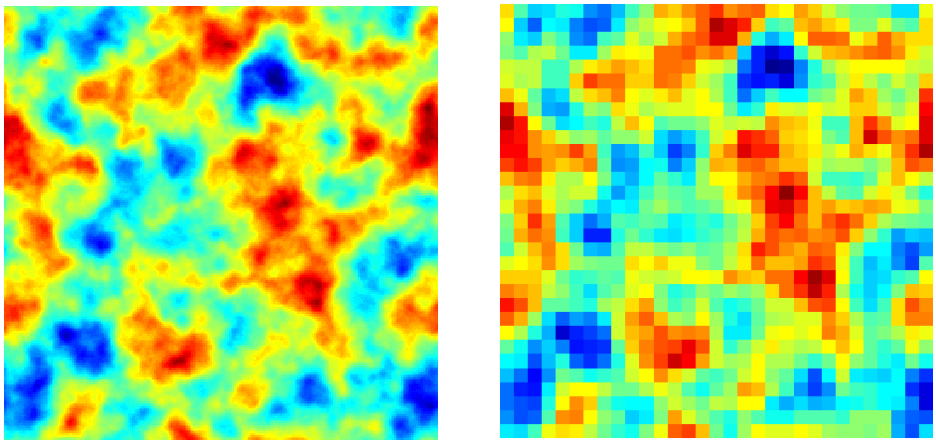


Figure 1-4: *Example of upscaling large quantities of fine scale permeability data to a coarse mesh suitable for simulation.*

in [89] by Wu, Efendiev and Hou where they consider grid block upscaled permeabilities $\tilde{K}$ for a periodic medium as the solution of

$$\tilde{K}\left(\frac{1}{|V|}\int_V \nabla p^\epsilon \;\; dx\right) = -\frac{1}{|V|}\int_V \mathbf{u}^\epsilon \;\; dx$$

where p^ϵ and $\mathbf{u}^\epsilon$ are the pressure and velocity solutions to

$$\mathbf{u}^\epsilon = -K^\epsilon \nabla p^\epsilon, \qquad \nabla \cdot \mathbf{u}^\epsilon = 0$$

in the block V with volume $|V|$ and subject to certain problem specific boundary conditions where K^ϵ is the fine scale permeability field. However, numerical upscaling methods (as well as other multiscale methods that split a coupled global problem into de-coupled local problems) have the problem that large errors result from the resonance between the finite element mesh scale H and the fine scales in the continuous problem ϵ, for example the H^1 and L_2 norms are of the order of ϵ/H which is comparatively large when H is of the same order as ϵ. This was shown in the error analysis in [89] which builds on the analysis for a multicale basis function method by Hou, Wu and Cai earlier in [50].

1.2.2.2 Multiscale basis functions

Instead of trying to upscale the high contrast coefficient to a coarser grid, replacing $\mathcal{A}$ by $\mathcal{A}_{\text{upscaled}}$, we can think about creating finite element basis functions that incorporate locally the fine scale features of the physical problem, so for example instead replace a piecewise linear basis function Φ by a multiscale one Φ^{MS}. This is what Hou and Wu did

in [49] to develop the multiscale finite element method for multi-dimensional problems with multiscale coefficients. The idea is to construct a multiscale basis function in each coarse grid element $\tau \in \mathcal{T}_H(\Omega)$ by solving a local homogeneous version of the governing equation

$$\nabla \cdot \mathcal{A}(x) \nabla \Phi_i^{\mathrm{MS}} = 0 \tag{1.15}$$

(in weak form) subject to what they term 'oscillatory' boundary conditions g_i found by solving the 1D problem

$$\frac{\partial}{\partial x} \mathcal{A}(x) \frac{\partial g_i(x)}{\partial x} = 0 \tag{1.16}$$

on each edge of τ where $g_i(x_j) = \delta_{ij}$ for the nodes x_j of τ (where $i, j = 1, ..., 4$ in [49] as they use rectangular elements). The small scales, now in the basis functions, then interact with the large scales through the variational formulation of the finite element method by Hughes et al [52] when solving the global finite element problem on the whole of the domain Ω (1.2) using V_H^{MS} as the span of these multiscale basis functions. Hou and Wu identified the importance of the choice of local boundary condition connecting the small scale bases to the macroscopic solution.

Similarly there is work by Jenny, Lee and Tchelepi [53] to construct a multiscale finite volume method that finds coarse scale transmissibilities, fluid flux across coarse element boundaries, using basis functions that incorporate the fine scale data. The difference between [53] and [49] is that Jenny et al work with a finite volume method which is a mass preserving discretisation unlike the standard FEM. They also solve a local version of the elliptic PDE subject to oscillatory boundary conditions but on the dual mesh which then allows them to construct boundary conditions for the local elliptic PDE on the original mesh.

1.2.2.3 Current convergence analysis without homogenisation

The recent work by Chu, Graham and Hou in [27] makes **no appeal to Homogenisation theory** to prove convergence estimates of their multiscale finite element method. Like in [49] they introduce the local homogeneous problem (1.15) to solve in order to obtain multiscale basis functions, but crucially obtain a boundary condition that ensures a priori that the finite element error is of first order in the energy norm and second order in the L_2 norm. They show that the 'oscillatory' boundary conditions (1.16) of [49] are in fact a special case of their local boundary conditions and the error estimates in [27] help to explain why many of the methods from homogenisation

techniques work very well also for interface problems without periodic coefficients. The method for constructing these multiscale basis functions is in fact quite simple, however the convergence analysis is very complicated. In Chapter 3 we seek to give an overview of the ideas from [27] to construct these multiscale basis functions and prove the convergence but also extend some results to more general settings.

1.2.2.4 Previous convergence analysis using homogenisation

The convergence analysis of the multiscale finite element method in [49] and in fact in most other works on multiscale FEMs is done by considering the periodic homogenisation problem

$$-\nabla \cdot \mathcal{A}_\epsilon \nabla u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega. \tag{1.17}$$

Here the coefficient $\mathcal{A}_\epsilon$ is of the form $\mathcal{A}_\epsilon(x) = \mathcal{A}(x/\epsilon)$ where $\epsilon > 0$ is a small parameter and $\mathcal{A}(y)$ is a smooth positive valued periodic function on the unit cell Y ($[0,1]^2$ in $\mathbb{R}^2$ for example). The analysis seeks to prove robust convergence with respect to the oscillation coefficient ϵ as in [49, 50] (i.e. the finite element error does not depend on ϵ as $\epsilon \to 0$) rather than robustness to the contrast in $\mathcal{A}(y)$ which is our focus here.

By homogenisation theory, the solution of (1.17) has an asymptotic expansion of the form

$$u = u_0(x) + \epsilon u_1(x, y) - \epsilon\theta_\epsilon + O(\epsilon^2) \tag{1.18}$$

where $y = x/\epsilon$ is the rapidly oscillating variable and $x \in \Omega$. In (1.18) u_0 is the solution of the leading order homogenised equation

$$-\nabla \cdot \mathcal{A}^*_\epsilon \nabla u_0 = f \quad \text{in } \Omega, \quad u_0 = 0 \quad \text{on } \partial\Omega \tag{1.19}$$

where $\mathcal{A}^*_\epsilon$ is the effective coefficient given by

$$(\mathcal{A}^*_\epsilon)_{ij} = \frac{1}{|Y|} \int_Y (\mathcal{A}_\epsilon)_{ik} (y) \left(\delta_{kj} - \frac{\partial}{\partial y_i} \chi^j \right) \quad dy \tag{1.20}$$

and χ^j is the periodic solution of the unit cell problem

$$\nabla_y \cdot \mathcal{A}(y) \nabla_y \chi^j = \frac{\partial}{\partial y_i} \mathcal{A}_{ij}(y) \tag{1.21}$$

with zero mean. This solution to the unit cell problem then gives the equation for the

first-order term $u_1(x, y)$ in (1.18) as

$$u_1(x, y) = -\chi^j \frac{\partial u_0}{\partial x_j}. \tag{1.22}$$

Normally u_1 is non-zero on the boundary $\partial\Omega$ and so the zero boundary condition is enforced through the first-order corrector θ_ϵ in (1.18). This is the solution of

$$\nabla \cdot \mathcal{A}_\epsilon(x)\nabla\theta_\epsilon = 0 \quad \text{in } \Omega, \quad \theta_\epsilon = u_1(x, x/\epsilon) \quad \text{on } \partial\Omega \tag{1.23}$$

Typical analysis of the standard FEM gives rise to an overly pessimistic error estimate in the H^1 norm that is $O(H/\epsilon)$ which is extremely poor unless $H << \epsilon$. Instead Hou and Wu showed, via the expansion (1.18), that for the multiscale FEM you get error estimates of the form

$$\|u - u_H\|_{H^1(\Omega)} \leq C_1 H \|f\|_{L_2(\Omega)} + C_2 \left(\frac{\epsilon}{H}\right)^{\frac{1}{2}} \tag{1.24}$$

for $\epsilon < H$ where C_1 and C_2 are independent of ϵ and H. Therefore the finite element error estimate is robust to the oscillation parameter ϵ as ϵ tends to zero. However they showed that there is a resonance effect in the first-order corrector θ_ϵ where a boundary layer of amplitude $O(1/\epsilon)$ exists when solving (1.23) with inexact boundary conditions on each element. Motivated by an example where $\mathcal{A}$ is separable ($\mathcal{A}(x, y) = \mathcal{A}_1(x)\mathcal{A}_2(y)$) Hou, Wu and Cai use the oscillatory boundary conditions (1.16) to remove this boundary layer (as $\theta_\epsilon = 0$) and consequently recommend using this technique for other coefficients $\mathcal{A}$ (see Section 6.2 [50]). They also propose an oversampling technique to remove it further, since the corrector θ_ϵ is only $O(1)$ away from the boundary they consider solving the homogenisation equations on a larger cell of size $\tilde{H} > H + \epsilon$. Numerical results show this method is very effective.

There are many ways to approach solving the Homogenisation problem (1.17). Early work on numerical homogenisation can be found in [19] by Bourgat where they examine the homogenisation problem with a periodic coefficient $\mathcal{A}_\epsilon$. Several error bounds with respect to the oscillation parameter ϵ are given as well as some early numerical experiments. Enquist and Runborg give a comprehensive overview of homogenisation techniques in [38] and introduce a multiscale finite element method for elliptic homogenisation problems. This is built upon by Henning and Ohlberger in [45] where they analyse a generalisation of this method to perforated domains and introduce some a posteriori error estimates. Other work of note is that of Allaire in [6] and along with Briane in [8] where they introduce the homogenisation problem for two scales and introduce tools for proving convergence properties of the homogenisation problem. Schwab

and Hoang build on this in [47] where they introduce the sparse tensor product finite element method for homogenisation problems on many scales thus in high dimensions, i.e. where

$$\mathcal{A}_\epsilon = \mathcal{A}(x, \frac{x}{\epsilon_1}, \frac{x}{\epsilon_2}, ..., \frac{x}{\epsilon_N}).$$

Here they also prove convergence results of their method with respect to the oscillation parameters ϵ_i. Recently there has been a lot of work by Owhadi and Zhang in [74] and [75] on upscaling methods for the homogenisation problem. They calculate solutions to the global harmonic problems

$$\begin{aligned} \nabla \cdot \mathcal{A} \nabla F_i &= 0 \qquad \text{in } \Omega \\ F_i(x) &= x_i \qquad \text{on } \partial\Omega \end{aligned}$$

to provide an N-dimensional map $F(x) = (F_1(x), ..., F_N(x))$ to transform a rapidly varying problem (e.g. homogenisation problems) into a smooth problem through the use of $(\nabla F)^{-1}$ where ∇F is the Jacobian given by $(\nabla F)_{ij} = \partial F_i / \partial x_j$. For example the problem: Find $u \in H_0^1(\Omega)$ such that

$$\nabla \cdot \mathcal{A} \nabla u = g \qquad \text{in } \Omega$$

would be transformed to: Find $u \in H_0^1(\Omega)$ such that

$$\nabla \cdot (\mathcal{A}(\nabla F)) \left((\nabla F)^{-1} \nabla u \right) = g \qquad \text{in } \Omega$$

where $(\nabla F)^{-1} \nabla u$ is now in $H^2(\Omega)$. This is shown in Figure 1-5 where $\nabla_F u = (\nabla F)^{-1} \nabla u$.

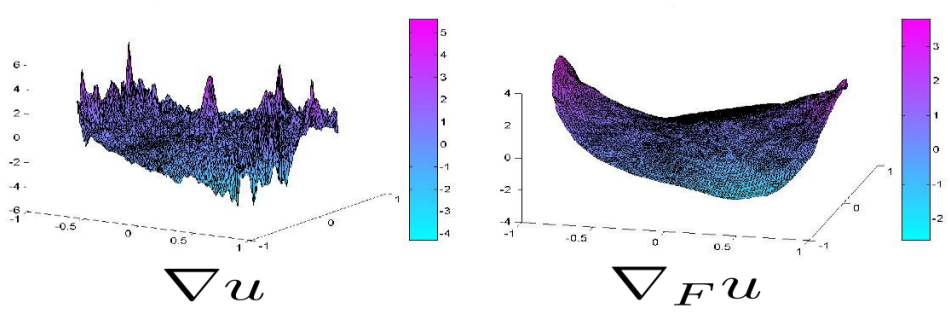


Figure 1-5: *Example of the Owhadi Zhang metric to obtain a smooth problem. [Courtesy of L. Zhang [91]]*

This has also had significant success with the multiscale interface problem without periodic coefficients but the metric based upscaling method has the drawback of requiring accurate global solves. Work is ongoing to reduce this to locally supported functions. It also provides some analysis of convergence properties for the homogenised problem.

1.2.2.5 Other multiscale methods outside the context of this thesis

There are several variants of the idea of multiscale basis functions that better approximate the solution to (1.1), i.e. have a smaller finite element error. The most popular of these is the Extended Finite Element Method (XFEM) by Belytschko, Dolbow and Moës in [69]. This is an implementation of the Generalised FEM by Babuška and Osborn and it seeks to enrich the approximation space by introducing additional basis functions that incorporate the non-smoothness of the solution. The standard XFEM solution takes the form

$$u_H(x) = \underbrace{\sum_{i\in\mathcal{N}} U_i\Phi_i(x)}_{\text{standard FE approximation}} + \underbrace{\sum_{i\in\mathcal{N}^*} a_i\Phi_i^*}_{\text{enrichment}} \tag{1.25}$$

where Φ_i is a standard set of finite element basis functions and U_i their corresponding weights for the set of mesh nodes $\mathcal{N}$. The enrichment is only done for a subset of these nodes $\mathcal{N}^* \subset \mathcal{N}$ with weights a_i and additional basis functions Φ_i^*. These enrichment functions take the form

$$\Phi_i^* = \phi_i^* \cdot \psi$$

where ϕ_i^* is another set of standard finite element functions (not necessarily the same as Φ_i) that specifically forms a partition of unity and ψ is a global enrichment function, chosen to incorporate the desired singularities, for example if there is a jump in the gradient of the solution (a weak singularity) then typically ψ is the signed distance function to the jump.

Belytschko et al were applying their method to crack propagation through a material and hence used basis functions that incorporated the discontinuous Heaviside function but still form a partition of unity. Like many of the other multiscale methods, XFEM has the advantage that a domain need only use uniform meshing rather than precisely resolving the crack. Consequently the mesh does not need updating as the crack propagates. The drawback of this method though is that the nature of the singularity needs to be known beforehand in order to know what type of global enrichment function ψ to equip the approximation space with. Also the introduction of the additional basis

functions Φ_i^* means that you have multiple degrees of freedom per node, if there are a lot of these additional basis functions then this can significantly increase the size of the linear system that needs to be solved and thus poses computational issues.

The XFEM approach is combined with the level set method in order to model moving interfaces in [82] by Sukumar, Chopp, Belytschko and Moës. The level set method removes the complication of tracking a moving interface because it replaces the interface by a function over the whole domain where the level set (where it is zero) describes the interface. Typically the level set takes the form

$$\gamma(x) = \begin{cases} < 0 & \text{if } x \text{ is inside the interface} \\ 0 & \text{if } x \text{ is on the interface} \\ > 0 & \text{if } x \text{ is outside the interface} \end{cases} \tag{1.26}$$

Then the entire function γ is updated to produce a new level set $\{x \in \Omega \mid \gamma(x) = 0\}$. Sukumar et al show how the method can be applied to interface problems with a weak discontinuity where there is a jump in the gradient of the solution.

Strouboulis, Babuška, Copps and Zhang also introduce the idea of creating additional enriched basis functions in [79, 81, 80] by solving a local problem around voids (considered to be the holes within a structure Ω_S where no material is present and given by $\mathbb{R}^2 \backslash \Omega_S$) and cracks (a split in the material with an infinitesimal gap between two connected edges) to get so called 'handbook functions' $\psi_j^{X_i}$ that numerically try to incorporate the nature of the singularity into these additional basis functions. The 'handbook space' is of dimension n_{hb} and thus $j = 1, ..., n_{\text{hb}}$. The local Neumann problem they solve is

$$\Delta \psi_j^{X_i} = 0 \qquad \text{in } \tilde{\omega}_{X_i}^{(1)} \tag{1.27}$$

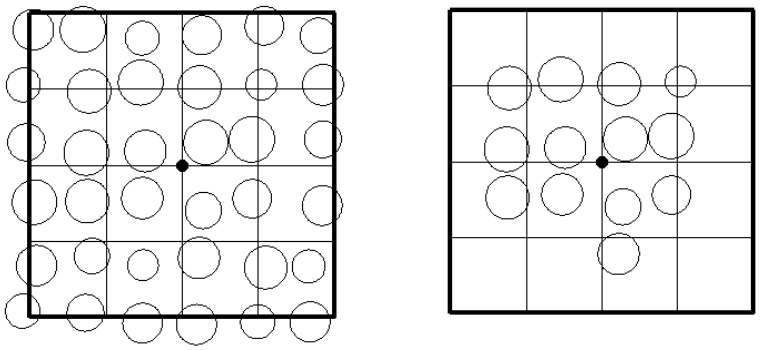
where $\omega_{X_i}^{(1)}$ is the set of elements connected to the node X_i and their neighbouring elements (Figure 1-6(a)) and $\tilde{\omega}_{X_i}^{(1)}$ is $\omega_{X_i}^{(1)}$ but with the voids not intersecting the neighbours of X_i removed (Figure 1-6(b)). If we let $z = x + iy \in \mathbb{C}$ for a point $\mathbf{x} = (x, y) \in \mathbb{R}^2$ and seek a handbook function of order p then the Neumann problem above is subject to the following boundary conditions

$$\frac{\partial}{\partial \mathbf{n}} \left(\psi_j^{X_i} \right) = \begin{cases} \nabla(\Re(z^p)) \cdot \mathbf{n} & \text{if } j \text{ is odd} \\ \nabla(\Im(z^p)) \cdot \mathbf{n} & \text{if } j \text{ is even} \end{cases} \qquad \text{on } \partial\omega_{X_i}^{(1)} \ , \tag{1.28}$$

i.e. $\partial\omega_{X_i}^{(1)}$ is the thick black line of Figure 1-6(a), and

$$\frac{\partial}{\partial n}\left(\psi_j^{X_i}\right) = 0 \qquad \text{on } \partial\tilde{\omega}_{X_i}^{(1)} \backslash \partial\omega_{X_i}^{(1)} \ , \tag{1.29}$$

where $\partial\tilde{\omega}_{X_i}^{(1)} \backslash \partial\omega_{X_i}^{(1)}$ is the boundaries of the voids left in $\tilde{\omega}_{X_i}^{(1)}$.



(a) $\partial\omega_{X_i}^{(1)}$. (b) $\partial\tilde{\omega}_{X_i}^{(1)}$.

Figure 1-6: *Examples of a local perforated domain $\partial\omega_{X_i}^{(1)}$ and then the restricted domain $\partial\tilde{\omega}_{X_i}^{(1)}$ used for the local handbook problem.*

Recently Mousavi, Grinspun and Sukumar in [70] have shown how to use solutions to the Laplace equation subject to zero Dirichlet conditions on the crack and zero Neumann conditions of the edge of the local enrichment domain to get the enrichment functions. Mousavi et al extend this to higher-order elements in [71]. This idea is like that of the multiscale finite element method in [27] in that it considers solving local problems to get basis functions that better approximate the solution. However it only deals with cracks and relies on a mixture of zero Dirichlet and Neumann boundary conditions for the local problems. It also uses these functions in addition to the original basis set whereas the multiscale FEM in [27] creates a set of basis functions with only one degree of freedom per node but still captures the behaviour of the interface problem. The multiscale method by Chu, Graham and Hou in [27] instead replaces the Φ_i in (1.25) with multiscale functions Φ_i^{MS} and does not have the enrichment part.

Another similar work that examines the idea of enrichment is that of Brezzi in [22, 21] with the residual-free bubble method. Here the enrichment functions in (1.25) are bubble functions meaning that for an element K of a mesh $\mathcal{T}_H(\Omega)$ they are functions in $H_0^1(K)$. It decomposes the approximate solution u_A into

$$u_A = u_H + u_B$$

where u_H is from the standard finite element space (e.g. the space of continuous piecewise linear functions) and u_B is from the space of functions whos restriction to each element is a bubble function. u_B in each triangle is the solution of the bubble equation; find $u_{B,K} \in H_0^1(K)$ such that

$$-\text{div}\left(\mathcal{A}\nabla u_{B,K}\right) = \text{div}\left(\mathcal{A}\nabla u_H\right) + f \qquad \text{in } K$$

where $u_{B,K} = u_B|_K$. This however does leave the problem of how to include variation along the element edges, Brezzi suggests the addition of edge functions to the decomposition.

The work closest to the multiscale finite element method in [27] is that of Li, Lin and Wu in [62] where they introduce the immersed finite element (IFE) method. Like in the multiscale finite element method, the immersed finite element method uses an unfitted mesh, i.e. the mesh does not have to line up with the interfaces. The IFEM then approximates the interface through each cut element as a straight line segment L_τ (Figure 1-7). By matching the jump condition

$$\mathcal{A}^- \frac{\partial^- \Phi_i^{\text{IFE}}}{\partial n} = \mathcal{A}^+ \frac{\partial^+ \Phi_i^{\text{IFE}}}{\partial n}$$

on the line segment L_τ (where $(\cdot)^+, (\cdot)^-$ represent a value taken from each side of the interface and n is the normal to L_τ) they created special basis functions on cut elements and consequently proved a first and second order convergence rate in the H^1 semi-norm and L_2 norm respectively. However, their error estimate was strongly dependent on the contrast (1.12) in the coefficient $\mathcal{A}(x)$. Chu, Graham and Hou showed that the immersed finite elements are in fact a special case of their multiscale basis functions when the interface intersects a coarse grid element as a straight line.

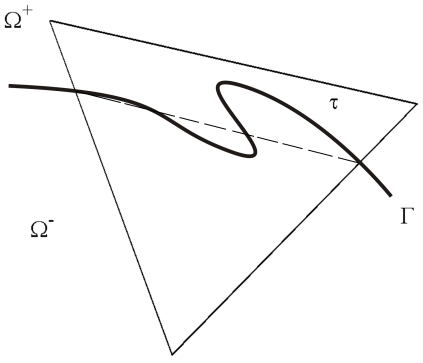


Figure 1-7: *Example of approximating an interface by a straight line through an element.*

There has also been a lot of work developing finite difference methods for the interface problem. One such method is the Immersed Boundary Method (IBM) by Peskin in [76] for elastic surfaces immersed in an incompressible viscous fluid and applied to biofluid dynamics problems. These problems involve complex domain geometries and immersed elastic membranes. Like the multiscale finite element method it employs a uniform Eulerian (fixed position with varying value) mesh $\mathcal{T}_H(\Omega)$ over a domain Ω. This describes the velocity field of the fluid and then it uses a Lagrangian (particle based) description of the membrane within the fluid. Unverdi and Tryggvason were also motivated by Peskin's method and developed a successful front tracking method for viscous incompressible multiphase flows in [83].

Another related finite difference work is the Immersed Interface Method (IIM) for elliptic interface problems and was developed by LeVeque and Li in [58]. The immersed interface method uses the jump condition across the interface to modify the finite difference approximation near the interface. When done properly this can achieve second order accuracy. The IIM can also be applied to the moving interface problem as in [48] by Hou et al and interface problems in irregular domains like in [32] by Dumett and Keener. Several extensions and improvements can be found in [5, 60, 61] by Li et al.

The Ghost Fluid Method (GFM) was developed by Fedkiw, Merriman, Aslam and Osher [39]. The GFM includes the jump condition in the finite difference discretisation in such a way that it can be implemented in an efficient way. The GFM has been applied to capture discontinuities in multimedium compressible flow [64] by Liu, Khoo and Wang and strong shock impacting problems [65] by Liu, Khoo and Yeo. It has been generalised to the elliptic interface problem in [66] by Liu, Fedkiw and Kang and its convergence property has been analysed in [67] by Liu and Sideris. Other related works include [26, 92] by Chern et al and Zhou et al respectively. There has actually been little progress in **coefficient robust** convergence results for finite difference methods for interface problems. In contrast, both [27] and this thesis provide coefficient robust a priori bounds for certain multiscale finite element methods.

1.2.3 Advances in adaptive multiscale methods

There have been several advances in constructing multiscale basis functions in a local fashion, however so far these have been under very specific assumptions that we will discuss in Chapter 3. While the multiscale basis functions given in [27] will satisfy these assumptions given sufficient local mesh refinement (i.e. the elements of the mesh $\mathcal{T}_H(\Omega)$ that do not satisfy the assumptions are divided until they do) it would be desirable to

have a method that can handle any irregular coefficient $\mathcal{A}(x)$ running though a unfitted mesh of elements (i.e. a mesh which does not necessarily match the discontinuities of $\mathcal{A}(x)$). In fact it would be useful to have a method that works for any high contrast second order elliptic problem of the form (1.1) rather than specifically the interface problem (1.11). This would then allow us to attack the linear elasticity problem and in fact this is discussed in more detail in Chapter 5. Future research may provide such a priori local boundary conditions for these more general settings but for now we consider trying to find these local artificial boundary conditions adaptively.

The procedures described here are different from conventional adaptive techniques that try to refine the mesh to have smaller elements in concentrated areas (h-refinement), increase the order of polynomial used for basis functions (p-refinement) or in fact move mesh nodes around to better approximate areas with more activity (r-refinement). Here, the idea is to adapt the shape of the basis function to better approximate the shape of the solution. This is like p-refinement but the basis functions are not necessarily polynomial. This is discussed further in Chapter 4.

1.2.3.1 Adaptive methods relevant to this thesis

The main work for adaptive basis function multiscale finite element methods came with the introduction of a local-global method by Durlofsky, Efendiev and Ginting in [36]. Here they proposed an adaptive local-global multiscale finite element method (ALG-MsFEM) that linked creation of the local multiscale basis functions to the global pressure solution. They proposed a two step method that used initial basis functions to approximate a global pressure solution, this solution is used to get a second more accurate set of basis functions to then obtain the final approximate solution from. They showed this ALG-MsFEM method to be effective for two-phase flow simulations. The method originated earlier in [24] and [23] by Chen et al where this local-global technique was used for upscaling the permeability values. This meant that in [24] and [23] the same approximation space is always used from one step to the next, the difference in [36] is that the local-global step updates the local boundary conditions and then the resulting approximation space is different from the starting space. We discuss the ALG-MsFEM further in Chapter 4 and look at how the method is far more versatile than stated in [36]. We examine improvements to the ALG-MsFE method and look at some of its properties. The oversampling technique involved in [36] uses linear boundary conditions on the oversampled local domain but this is not ideal. Chu, Efendiev, Ginting and Hou showed in [28] that using actual boundary conditions from the two-phase flow problem gave much better accuracy and we will also show in this

thesis that use of the 'oscillatory boundary conditions' (1.16) from [49] also gives a much better result in the context of high contrast interface problems.

1.2.3.2 Historical adaptive multiscale methods

There has been a lot of recent work on this subject but particularly in the field of reservoir modelling in porous media flow. As techniques for providing geological data have improved so has the size of the data sets available. The difficulty is in including all of this fine scale information into a model of the reservoir (as was discussed earlier regarding upscaling techniques in Section 1.2.2 and shown in Figure 1-4). Aarnes started addressing this issue in [1] and along with Kippe and Lie in [3]. They raised the point that the fine scale structures have a non-trivial impact on the global flow solution. In [3] they demonstrated this by considering two types of local boundary condition to get the multiscale basis functions. The first used only local information while the second included information obtained from an initial approximation to the global velocity field. The results showed that the oil production curves better matched the fine scale solution when these so-called 'global boundary conditions' were used. It is worth noting that this work is all for the mixed form of the two-phase flow problem and Aarnes along with Krogstad and Lie introduce adaptivity in [4] in the form of hierarchical mesh refinement of the non-uniform coarse mesh involved. This was extended further in [2] by Aarnes and Efendiev where the multiscale basis functions were replaced in areas with sharp fronts by a solution to a local transport equation.

1.2.3.3 Other adaptive multiscale method literature

Another adaptive multiscale method for solving (1.11) is given by Nolen, Papanicolaou and Pironneau in [72]. Here they develop a framework for creating an approximate solution via projections on to spaces capturing the coarse and the fine details, i.e. for $u \in H_0^1(\Omega)$ it will have a decomposition

$$u = \mathcal{P}_C u + (\mathcal{I} - \mathcal{P}_C)\, u = \text{coarse approximation} + \text{details} \tag{1.30}$$

where $\mathcal{P}_C$ is a projection on to a finite dimensional approximation space X_C. Their method involves finding the map $\mathcal{M} : \nabla X_C \to X_F$ to reconstruct the fine details, where X_F is the image of $(\mathcal{I} - \mathcal{P}_C)$, such that

$$u = u_C + \mathcal{M}\left(\nabla u_C\right). \tag{1.31}$$

The map $\mathcal{M}$ is found by solving

$$\int_\Omega \mathcal{A}(\mathcal{I} + \nabla\mathcal{M})\nabla u_C \cdot \nabla v = \int_\Omega fv \qquad \text{for any } v \in X_C \tag{1.32}$$

where $\mathcal{M}$ is decomposed as the operator $\mathcal{M}(\nabla v) = \mu_F + \mathcal{M}_0(\nabla v)$. $\mathcal{M}_0$ and μ_F are then found by solving

$$\int_\Omega \mathcal{A}\nabla(\nabla\mathcal{M}_0\nabla w) \cdot \nabla v = -\int_\Omega \mathcal{A}\nabla w \cdot \nabla v \qquad \text{for any } v \in X_F, \quad w \in X_C \tag{1.33}$$

and

$$\int_\Omega \mathcal{A}\nabla u_F \cdot \nabla v = \int_\Omega fv \qquad \text{for any } v \in X_F. \tag{1.34}$$

If only the coarse scale component is being computed then μ_F is not needed. They approximate new basis functions $w_k = \phi_k^C + \mathcal{M}_0(\nabla\phi_k^C)$ where $\mathcal{M}_0$ applied to a basis function ϕ_k^C of X_C, however $\mathcal{M}_0$ is nonlocal. Therefore Nolen et al approximate $\mathcal{M}_0(\nabla\phi_k^C)$ locally and improve it by using an oversampling technique to capture a more accurate projection of a starting basis function; the adaptivity of their method comes from determining how large the oversampling region should be to achieve a good approximation to $\mathcal{M}_0(\nabla\phi_k^C)$.

While the work in [72] provides a very general framework it does not provide results that help with proving coefficient robust finite element error estimates. There is freedom to choose the projection $\mathcal{P}$ in (1.30) and they choose the H_0^1 orthogonal projection. Because of the quasi-optimality result

$$|u - u_H|_{H_0^1} \leq \frac{\max \mathcal{A}(x)}{\min \mathcal{A}(x)} \inf_{v \in X_C} |u - v|_{H_0^1} = \frac{\max \mathcal{A}(x)}{\min \mathcal{A}(x)} |u - \mathcal{P}u|_{H_0^1} \ .$$

This suggests that the H^1 error for $u - \mathcal{P}u$ is smaller than the finite element error when the contrast (1.12) is large. In fact using the orthogonal projection with respect to the inner product $(u, v)_{\mathcal{P}_C} = (\mathcal{A}\nabla u, \nabla v)_{L_2}$ (the Galerkin solution) we actually get optimality

$$|u - u_H|_{H_0^1,\mathcal{A}} = (\mathcal{A}\nabla(u - u_H), \nabla(u - u_H))_{L_2}^{\frac{1}{2}} \leq \inf_{v \in X_C} |u - v|_{H_0^1,\mathcal{A}} = |u - \mathcal{P}u|_{H_0^1,\mathcal{A}} \ .$$

As we will show in Chapters 2 and 3 we can use this to get coefficient independent relative error estimates. Nolen et al mention using this inner product and how it results in $\mathcal{M}_0(\nabla u) \equiv 0$ and thus all the fine scales are encompassed in μ_F. With the H_0^1 orthogonal projection the fine scale information is incorporated into the basis

function ϕ_k^C by calculating the new function $w_k = \phi_k^C + \mathcal{M}_0(\nabla\phi_k^C)$. However they do not discuss the implications of changing the coarse space X_C instead. This is the idea of the multiscale methods in [36], [27] and this thesis instead of approximating the operator $\mathcal{M}_0$. Using the idea of finding coarse level multiscale basis functions it is possible to prove explicit coefficient robust finite element error estimates.

1.2.4 Application of multiscale methods to structural optimization

An interesting point to note about the multiscale finite element methods mentioned above is that they have only been applied to flow problems. The methods above, particularly the ALG-MsFEM in [36], can be stated in a very general way to cover other engineering problems. In this thesis we also present a generalisation of the method in [36] to problems in linear elasticity and specifically to the area of structural/topology optimisation. This field seems to have developed independently within mechanical engineering with very little cross over into multiscale modelling even though the problems have a number of similarities.

1.2.4.1 Current fixed mesh structural optimisation methods

Allaire, Jouve and Toader introduced the idea of using an 'ersatz' material to extend a structure to the whole of a design domain in [9]. This effectively fills in voids with a 'ghost' material that mimics voids but avoids the finite element stiffness matrix being singular by utilising a weak material with a small but non-zero Young's modulus in the voids. This makes it simpler than the immersed interface method to implement. This formulation allows an area weighted approach (where for two basis functions ϕ_i, ϕ_j the stiffness matrix values are given by $A(\phi_i,\phi_j)|_\tau = \nabla\phi_i \cdot \left(\int_\tau \mathcal{A}\right)\nabla\phi_j \approx \nabla\int_\tau \phi_i \cdot \mathcal{A}\nabla\phi_j$, thus $\int_\tau \mathcal{A}$ is effectively an area weighting and is exact for linear basis functions) to be considered for solving the linear elasticity problem and was first proposed by García-Ruíz and Steven in [40]. In their analysis they also showed that most of the error occurs at the boundaries of the structure just as is the case for the second order elliptic interface problem. There are also errors that arise in the fixed grid method that are mesh fit dependent, this means that a large error can occur if the mesh almost resolves the interface in one element but then is a very poor fit in a neighbouring element. Some of these issues are addressed in [33] and [34] by Dunning, Kim and Mullineux with the introduction of isoparametric elements. A lot of these isoparametric methods are very similar to the idea of immersed finite elements in [62] that try to approximate the interface by a straight line. For this reason it would be of interest to formulate a

general version of the immersed finite element method to apply to these linear elasticity problems. Going further than this, we know that the multiscale finite elements of [27] are a generalisation of these immersed elements and so it would be of more interest to formulate a linear elasticity version of this method. We will present this as well as a general version of the adaptive method in [36] applied to linear elasticity for use in structural optimisation.

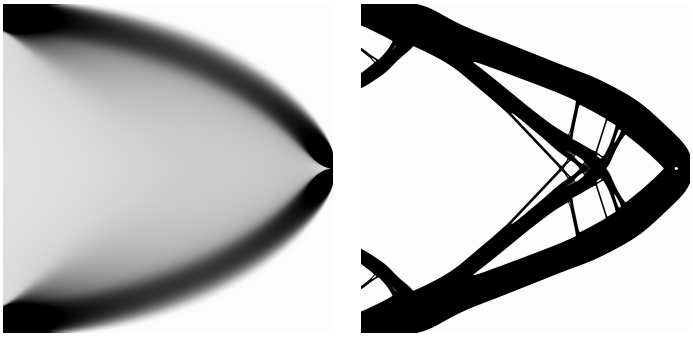
1.2.4.2 Historical development of structural optimisation methods

Structural optimisation tries to find the best configuration of a limited amount of material to do a task, as was discussed in the applications part of Section 1.1. For example it tries to find the best 2D cantilever, a structure that is attached on one side to a wall and with a load hanging from the other side (Figure 1-2(c)). Normally a solid bar would be the best structure for this but the idea is to reduce the amount of material used. This is usually done by introduction of holes (known as microstructure) but then the question is over what size and shape they should take. The idea is to take an initial guess and then update the configuration of material but this introduces the problem of moving interfaces and changing topology. To deal with this Osher and Sethian introduced the level set method in [73] somewhat akin to the level set description used by XFEM in Section 1.2.2. Here they describe the equations of motion deriving from Hamilton-Jacobi formulations. That is, the front ϕ propagates with a speed F dependent on the curvature K at that point according to

$$\frac{\partial \phi}{\partial t} - F(K)\,|\nabla \phi| = 0. \tag{1.35}$$

They applied their numerical algorithms to crystal growth and flame propagation. Then later in 2000 Sethian and Wiegmann applied this to structural boundary design in [78] where they determine the velocity of the boundary of a structure (analogous to the front ϕ) by the stresses on them. The key point about the level set method is that it separates the optimisation process from the linear elasticity problem. At each step the linear elasticity problem is solved and then the result from this is used to update the level set. Interestingly and to the best of our knowledge [78] is the only paper to make use of a multiscale finite difference method of the type used in flow problems mentioned above. They utilise the explicit jump immersed interface method from [88] by Wiegmann and Bube which is a generalisation of the immersed interface method by LeVeque and Li in [58]. The numerical algorithm for describing the update process in structural optimisation is given by Wang, Wang and Guo in [85].

The reason that these level set methods have become popular is that the previous work on using homogenisation techniques to solve the linear elasticity problem (e.g. [7] by Allaire et al) proved unsatisfactory. While homogenisation replaces the more difficult problem of where to locate material with the easier problem of what density of composite to use, it creates structures that are unrealistic to construct as arbitrary densities and arbitrarily small scales cannot be manufactured. Figure 1-8(a) shows the homogenised solution for a 2D cantilever where the grey scale shows varying density. This problem can be overcome by incorporating penalty functions into the homogenisation process (Figure 1-8(b)) but these are very specific to the situation. Instead the level set method provides a very easy way to define a shape where there is material or a void and no densities in between.



(a) Homogenisation solution. (b) Homogenisation solution with a penalty.

Figure 1-8: *Examples of the solution to a structural optimisation problem with varying density and then binary material (Courtesy of P.A. Browne).*

1.3 The main achievements of the thesis

- Proving that the standard finite element error is $O(H^{\frac{1}{2}-\epsilon})$ in general for the energy norm, crucially with a constant independent of the contrast.
- Proving that an $O(H)$ convergence rate in the energy norm independent of the contrast can be restored with sufficient mesh refinement near interfaces.
- Extending to contrast independent finite element errors in the L_2 norm with corresponding $O(H^{1-2\epsilon})$ convergence rate in general and $O(H^2)$ with sufficient refinement near the interfaces.
- Proving a relative bound for the finite element error of the multiscale method.

- Extending the proof of the regularity theory for the multiscale method to multiple inclusions.
- A new generalisation of the adaptive multiscale method from Durlofsky, Efendiev and Ginting.
- An extension to their method to give a conforming method that still has the superior convergence of the "non-conforming" (EDG2) method. The enhanced adaptive method is shown to have a convergence rate of $O(H^2)$ in the L_2 norm even when the mesh does not resolve the jumps in the coefficient and when the coefficient contains corner point singularities or boundary layers.
- A novel application of the adaptive multiscale method to linear elasticity specifically examining structural optimisation.
- The creation of a parallel version of the adaptive multiscale method and a scaling analysis of it.
- Substantial numerical implementations relevant to general heterogeneous media.

1.4 The structure of the thesis

- Chapter 2 gives a detailed description of the high contrast elliptic problem, and also proves a new error bound for the finite element approximation that gives the dependence on the contrast explicitly.

- Chapter 3 describes the multiscale finite element method from [27] and extends the result to give a relative bound for multiple inclusions.

- Chapter 4 describes an adaptive multiscale method that has its origins in [36] but is described here in the normal finite element setting. The chapter gives a much more general description of the method and introduces a new change to the method to retain the power of the non-conforming method but makes it conforming. Numerical convergence of the method is also examined as well as showing its power when the solution contains singularities.

- Chapter 5 describes how this new adaptive method can be applied to the linear elasticity problem specifically to help in the field of structural optimisation. It

describes what structural optimisation is and the problems it presents, it gives a mathematical definition of the problem as well as showing how the adaptive multiscale method can be applied to it. Several benchmark problems are examined to show the strength of the new method.

- Chapter 6 describes how the method can be made practical by performing it in parallel. The parallelisation of the method is set out in detail and scaling studies are performed to show how successful the method is for scaling to many processors.

- Chapter 7 draws the thesis to an end with conclusions and suggestions for further work.

2

A priori error estimates for elliptic interface problems with high contrast

2.1 Problem definition

In this chapter we introduce the elliptic PDE with high contrast heterogeneous coefficient in more detail. This depends on a coefficient field $\mathcal{A}(x)$ that could be rapidly varying on a small scale. In general this is too difficult to obtain rigorous theoretical results for, so to make this more tractable we introduce the simpler high contrast interface problem where the domain contains inclusions. Now instead the coefficient field $\mathcal{A}(x)$ is slowly varying in each of those inclusions but may jump across the interface between inclusions. The "multiscale" property of this simpler problem arises from the geometry of the inclusions and the contrast in the coefficient field (the ratio between the maximum and minimum of $\mathcal{A}(x)$) which may be unbounded.

The main result in this chapter is a proof of a priori finite element error estimates for the interface problem that are, crucially, independent of the contrast in the coefficient field $\mathcal{A}(x)$ and do not require the mesh to resolve the coefficient jumps. This is a new result. The details of the proof are quite technical but we seek to step through the ideas of the proof in an accessible way to reach the final estimates.

The chapter will proceed as follows. We will start with a clear and detailed description of the general high contrast elliptic PDE and its simplification to the high contrast interface problem. We will then describe the finite element method and introduce corresponding notation. This will give enough tools to describe the idea of the proof of the contrast independent finite element error estimates. Following that we will describe how the proof proceeds for a single element that is cut by an interface, which will help to clarify the argument and make it easier to follow. Finally we will combine all of these ideas and results together to obtain the finite element error estimate on the whole domain. We start by giving a definition of the spaces we will be using.

Definition 2.1. *In this thesis a domain is defined to be a bounded open set in* $\mathbb{R}^2$. *For a domain* $\Omega \subset \mathbb{R}^n$ *and a function* $v : \Omega \to \mathbb{R}$ *define the* L_2 *and* H^1 *norms respectively by*

$$\|v\|_{L_2(\Omega)} = \left(\int_\Omega |v|^2 \quad dx \right)^{\frac{1}{2}} , \qquad \|v\|_{H^1(\Omega)} = \left(\int_\Omega |v|^2 + |\nabla v|^2 \quad dx \right)^{\frac{1}{2}}$$

and the H^1 *seminorm by*

$$|v|_{H^1(\Omega)} = \left(\int_\Omega |\nabla v|^2 \quad dx \right)^{\frac{1}{2}} .$$

We will also make use of the fractional order Slobodeckiĭ seminorm [p74 McLean [68]] defined by

$$|v|_{H_0^{1+\epsilon}(\Omega)} = \int_\Omega \int_\Omega \frac{|Dv(x) - Dv(y)|^2}{|x-y|^{2+2\epsilon}} dx \quad dy$$

Then define the spaces

$$\begin{aligned} L_2(\Omega) &= \left\{ v : \Omega \to \mathbb{R} \mid \|v\|_{L_2(\Omega)} < \infty \right\} , \\ H^1(\Omega) &= \left\{ v \in L_2(\Omega) \mid \|v\|_{H^1(\Omega)} < \infty \right\} , \\ H_0^1(\Omega) &= \overline{C_0^\infty(\Omega) \cap H^1(\Omega)} , \end{aligned}$$

where $C_0^\infty(\Omega)$ *is the set of infinitely differentiable continuous functions with non-zero support only on some part of* Ω *and* $\bar{\cdot}$ *denotes the closure of a set.*

In Chapter 1 we introduced the general elliptic PDE (1.1) that depends on a heterogeneous coefficient $\mathcal{A}(x)$. In order to produce theoretical a priori error estimates, throughout Chapters 2 and 3, we will restrict to the high contrast elliptic interface problem below. Note that this is only to obtain theoretical results and the adaptive multiscale finite element method in Chapter 4 will be applicable to the general elliptic PDE (1.1).

Problem 2.2. *(The Variational Interface Problem) Find* $u \in H_0^1(\Omega)$ *such that*

$$a_\Omega(u, v) = L_\Omega(v) \text{ for any } v \in H_0^1(\Omega) . \tag{2.1}$$

Let $a_\Omega(\cdot, \cdot)$ *be the bounded and coercive bilinear form*

$$a_\Omega(u, v) = \int_\Omega \nabla u \cdot \alpha \nabla v \quad dx \tag{2.2}$$

where $u, v \in H^1(\Omega)$ with a scalar piecewise constant permeability field $\alpha(x) \geq 1$ for any $x \in \Omega$. Let $L_\Omega(\cdot)$ be a functional of the form

$$L_\Omega(v) = \int_\Omega fv \quad dx \tag{2.3}$$

on a bounded polygonal domain $\Omega \subset \mathbb{R}^2$ and $v \in H^1(\Omega)$. We also assume $f \in L_2(\Omega)$.

The notion of a bounded and coercive bilinear form also introduces the energy norm.

Definition 2.3. *Given a domain $\Omega \subset \mathbb{R}^n$ and the bilinear form $a_\Omega(\cdot,\cdot)$ in (2.2) then the energy norm is defined for a function $u \in H^1(\Omega)$ as*

$$|u|_{H^1(\Omega),\alpha} = a_\Omega(u,u)^{\frac{1}{2}} \ . \tag{2.4}$$

In order to specify the interface problem (Problem 2.2) more precisely we need to define the coefficient α in more detail. In this thesis a domain is defined to be a bounded open set in $\mathbb{R}^2$.

Definition 2.4. *Let $\Omega \subset \mathbb{R}^2$, be a domain with a smooth or polygonal boundary $\partial\Omega$. Suppose Ω contains a finite number of* **inclusions** *denoted $\Omega_1, ..., \Omega_m$ where each Ω_i is the closure of a domain in Ω with smooth boundary and the inclusions are disjoint (i.e. if $i \neq j$ then $\Omega_i \cap \Omega_j = \emptyset$ for $i, j = 1, ..., m$). Consequently let $\Omega_0 = \overline{\left(\Omega \backslash \bigcup_{i=1}^{m} \Omega_i\right)}$ be the* **background inclusion**.

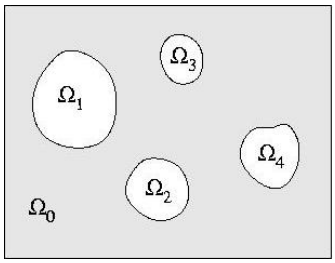


Figure 2-1: *An example Ω domain with inclusions.*

The analysis later will also require the notion of interfaces between inclusions, so this is defined in the following definition.

Definition 2.5. *For each inclusion Ω_i $i = 1, ..., m$, we define the interface between Ω_i and Ω_0 as*

$$\Gamma_i = \Omega_i \cap \Omega_0 \ . \tag{2.5}$$

In Problem 2.2 we stated that the coefficient α was piecewise constant. Now that we have the definition of inclusions we can define this notion precisely.

Assumption 2.6. *We assume that*

$$\alpha(x) = \alpha_i \quad \textit{for all } x \in \Omega_i \tag{2.6}$$

where $\alpha_i \geq 1$ is a constant on each Ω_i.

Significantly, it is the permeability field $\alpha(x)$ that introduces the contrast into the problem in the following definition.

Definition 2.7. *Suppose $\alpha(x)$ is a permeability field on a domain Ω. Then the ratio between the maximum and minimum of $\alpha(x)$,*

$$\max_{x\in\Omega} \alpha(x) \ / \ \min_{x\in\Omega} \alpha(x) \tag{2.7}$$

is known as the **contrast** *and in some applications (e.g. porous media flow) can be very large, often several orders of magnitude. When this ratio is large it is known as a* **high contrast** *field.*

As we are interested in the situations when the contrast (2.7) becomes very large we simplify matters by considering these two important cases:

$$\textbf{CASE I: } \alpha_0 = 1 \qquad \text{and} \qquad \hat{\alpha} := \min_{i=1,\dots m} \alpha_i \to \infty \ , \tag{2.8}$$

$$\textbf{CASE II: } \hat{\alpha} := \alpha_0 \to \infty \text{ and } \qquad \max_{i=1,\dots m} \alpha_i \leq K, \tag{2.9}$$

for some bounded positive constant $K > 0$ where $\hat{\alpha}$ represents a large "contrast parameter". Case I considers the scenario where the coefficients in the island inclusions becomes large compared to the background inclusion, whereas Case II is the opposite considering when the coefficient in the background inclusion becomes large relative to the island inclusions. The aim of this chapter is then to prove a robust finite element error bound of the form

$$|u - u_H|_{H^1(\Omega),\alpha} \leq C H^{\frac{1}{2}-\epsilon} \left(H^{\frac{1}{2}+\epsilon} + \delta_H^{\frac{1}{2}-\epsilon} \right) \|f\|_{L_2(\Omega)} \ , \tag{2.10}$$

for arbitrary $\epsilon > 0$, where u_H is the solution to the finite element problem (1.2) (for more detail see Section 2.2.1), C is independent of the contrast and δ_H is the ratio of the size of elements near the interface and H. So if $\delta_H \sim H$ then elements near the interface have size about H^2 and the finite element method converges with almost optimal order $O(H^{1-2\epsilon})$ independent of the discontinuity in $\alpha(x)$. For uniform meshes

the rate of convergence in (2.10) is $O(H^{\frac{1}{2}-\epsilon})$ **independent** of the contrast. We will also present a corresponding L_2 error estimate of the form

$$\|u - u_H\|_{L_2(\Omega)} \leq CH^{1-2\epsilon}\left(H^{\frac{1}{2}+\epsilon} + \delta_H^{\frac{1}{2}-\epsilon}\right)^2 \|f\|_{L_2(\Omega)} \ . \tag{2.11}$$

In this thesis we are mainly concerned with the robustness of this error estimate. Robustness means that the constant C in the above equation does not depend on $\hat{\alpha}$ as $\hat{\alpha}$ tends to infinity. This aim allows us to motivate the reasoning for taking a permeability field $\alpha(x)$ such that $\alpha \geq 1$ in the following remark.

Remark 2.8. *The restriction in Problem 2.2 to a permeability field $\alpha(x)$ where*

$$\alpha(x) \geq 1 \qquad \text{for any } x \in \Omega$$

can be relaxed. Suppose instead we want to solve the problem

$$\begin{cases} \int_\Omega \nabla u \cdot \mathcal{A}\nabla v \ \ dx = \int_\Omega fv \ \ dx & \text{for any } v \in H_0^1(\Omega) \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{2.12}$$

for $u \in H_0^1(\Omega)$ where the permeability field $\mathcal{A}(x)$ may approach zero on one or more inclusions. If we introduce a scaling

$$\alpha(x) = \frac{\mathcal{A}(x)}{\mathcal{A}_{min}}$$

where $\mathcal{A}_{min} = \min_{x\in\Omega} \mathcal{A}(x)$ and then (2.12) becomes an interface problem of the form in Problem 2.1 where

$$\int_\Omega \nabla u \cdot \alpha\nabla v \ \ dx = \int_\Omega \nabla u \cdot \frac{\mathcal{A}(x)}{\mathcal{A}_{min}}\nabla v \ \ dx = \int_\Omega \frac{f}{\mathcal{A}_{min}} v \ \ dx \ .$$

So from (2.10) we obtain the error estimate

$$|u - u_H|_{H^1(\Omega),\mathcal{A}/\mathcal{A}_{min}} \leq \frac{CH^{\frac{1}{2}-\epsilon}}{\mathcal{A}_{min}} \|f\|_{L_2(\Omega)}$$

which in the energy norm corresponding to $\mathcal{A}(x)$ gives

$$|u - u_H|_{H^1(\Omega),\mathcal{A}} \leq \frac{CH^{\frac{1}{2}-\epsilon}}{\mathcal{A}_{min}^{\frac{1}{2}}} \|f\|_{L_2(\Omega)} \, .$$

Thus we have the error estimate

$$\frac{|u - u_H|_{H^1(\Omega),\mathcal{A}}}{\mathcal{A}_{min}^{-\frac{1}{2}}} \leq CH^{\frac{1}{2}-\epsilon} \|f\|_{L_2(\Omega)} \ . \tag{2.13}$$

with C independent of the maximum value of the rescaled coefficient α. What we will show in Chapter 3 is that under suitable conditions the solution itself tends to infinity as $\mathcal{A}$ tends to zero with the bound

$$C(f)\mathcal{A}_{min}^{-\frac{1}{2}} \leq \max_i |u|_{H^2(\Omega_i)} \tag{2.14}$$

where $C(f)$ depends on f and thus (2.13) implies a relative error estimate of the form

$$\frac{|u - u_H|_{H^1(\Omega),\mathcal{A}}}{\max_i |u|_{H^2(\Omega_i)}} \leq C(f) H^{\frac{1}{2}-\epsilon} ,$$

showing robustness of the finite element error as $\mathcal{A}/\mathcal{A}_{min}$ tends to infinity.

2.2 Robustness of the standard finite element method

2.2.1 The finite element problem

For Problem 2.2 we shall show that the finite element error $|u - u_H|_{H^1(\Omega),\alpha}$ is independent of the contrast parameter $\hat{\alpha}$ (see (2.8) and (2.9)) for the coefficient function $\alpha(x)$. To set up the finite element problem we first introduce the concept of a mesh on the domain Ω.

Definition 2.9. *Given a polygonal domain $\Omega \subset \mathbb{R}^2$, let $\mathcal{T}_H(\Omega)$ be a subdivision of Ω into closed triangles with the properties that:*

1. *$\bar{\Omega} = \bigcup\{\tau : \tau \in \mathcal{T}_H(\Omega)\}$ and the elements have pairwise disjoint interiors.*
2. *If τ_1, $\tau_2 \in \mathcal{T}_H(\Omega)$ and $\tau_1 \neq \tau_2$ then $\tau_1 \cap \tau_2$ is either empty or an edge or a vertex.*

Then we apply the finite element method to approximate the high contrast interface problem (Problem 2.2) as described in Section 1.1 and (1.2). To do this we construct a set of basis functions whos span forms a finite dimensional subspace V_H of $H_0^1(\Omega)$.

Definition 2.10. *Let V_H be a finite dimensional subspace of $H_0^1(\Omega)$. In particular let $V_H^{\mathbb{P}_1}$ be the space of continuous functions that are linear on each element of the mesh $\mathcal{T}_H(\Omega)$, i.e.*

$$V_H^{\mathbb{P}_1} = \left\{ v \in C^0(\Omega) \mid \quad v|_\tau \in \mathbb{P}_1 \text{ for any } \tau \in \mathcal{T}_H(\Omega) \text{ and } v = 0 \text{ on } \partial\Omega \right\} \tag{2.15}$$

where $\mathbb{P}_1$ is the space of linear polynomials.

Then we look for an approximate solution $u_H \in V_H$ by solving (2.1) in $V_H^{\mathbb{P}_1}$.

Problem 2.11. *(The Finite Element Interface Problem) Find $u_H \in V_H$ such that*

$$a_\Omega(u_H, v_H) = \int_\Omega \nabla u_H \cdot \alpha \nabla v_H \, dx = \int_\Omega f v_H \, dx = L_\Omega(v_H) \quad \text{for all } v_H \in V_H \, . \tag{2.16}$$

Note that in particular we will take $V_H = V_H^{\mathbb{P}_1}$ in this chapter and refer to the finite element problem using this $V_H^{\mathbb{P}_1}$ as the **standard finite element problem**. In later chapters we will look at constructing better multiscale approximation spaces V_H^{MS}. For this thesis we will also have to define some commonly used notation to simplify the proofs and descriptions.

Notation 2.12. *For a domain $\sigma \subset \mathbb{R}^2$, define H_σ as the diameter of σ. So if σ is a triangle, H_σ is the length of the longest side. Also define ρ_σ as the diameter of the largest inscribed ball in σ.*

Notation 2.13. *For the mesh $\mathcal{T}_H(\Omega)$, the mesh diameter H is defined as*

$$H = \max_{\tau \in \mathcal{T}_H(\Omega)} H_\tau \tag{2.17}$$

and the mesh $\mathcal{T}_H(\Omega)$ becomes finer as H tends to zero.

It is also important to introduce some notation to make reading the proofs easier by removing insignificant constants.

Notation 2.14. *$g_1 \lesssim g_2$ means that there exists a constant $C > 0$, independent of the solution u, the load function f, the permeability field α and the mesh diameter H such that $g_1 \leq C g_2$. Also $g_1 \sim g_2$ means $g_1 \lesssim g_2$ and $g_2 \lesssim g_1$.*

Particularly, this notation is used when the hidden constant is independent of the mesh and the coefficient $\alpha(x)$. The dependence of any error bound on the coefficient function $\alpha(x)$ will be explicitly stated. For simplicity we assume shape regularity of the mesh $\mathcal{T}_H(\Omega)$, defined as follows.

Assumption 2.15. *We will assume that the mesh $\mathcal{T}_H(\Omega)$ is* **shape regular**, *i.e.*

$$1 \leq \max_{\tau \in \mathcal{T}_H(\Omega)} \frac{H_\tau}{\rho_\tau} \leq C \tag{2.18}$$

for some bounded $C \geq 0$. Note that error estimates will depend on this C and the lower bound follows from the fact that $\rho_\tau \leq H_\tau$.

It is necessary to label the nodes of the mesh and identify those nodes that lie on the boundary of the domain and those that lie away from the boundary.

Notation 2.16. *Let $\mathcal{N}(\mathcal{T}_H(\Omega))$ be the set of nodes of elements in the mesh $\mathcal{T}_H(\Omega)$. Also define $\mathcal{N}_0(\mathcal{T}_H(\Omega))$ as the set of nodes on the interior of Ω and $\mathcal{N}_D(\mathcal{T}_H(\Omega))$ as the set of nodes on the boundary of $\partial\Omega$.*

To help define elements cut by the interface we also need notation for the interior of a closed set.

Notation 2.17. *For a closed set τ, τ^o means the interior of τ (i.e. $\tau^o = \tau \backslash \partial\tau$).*

We are particularly concerned with the case when the interfaces Γ_i in Definition 2.5 run through the inside of elements and thus a linear finite element can not approximate the jump in gradient of the solution u very accurately. If the interface only intersects the boundary of elements then the finite element mesh $\mathcal{T}_H(\Omega)$ is said to **resolve** the interface. We will see that in this case usual error estimates from finite elements apply. The set of cut elements is defined as follows.

Definition 2.18. *The set of cut elements $\mathcal{T}_H^C(\Omega) \subset \mathcal{T}_H(\Omega)$ is given by*

$$\mathcal{T}_H^C(\Omega) = \{\, \tau \in \mathcal{T}_H(\Omega) |\, \tau^o \cap \Gamma_i \neq \emptyset \text{ for some } i = 1, ..., m \} \ . \tag{2.19}$$

As well as the cut elements themselves we also need a definition of the elements next to them. These will be known as border elements, defined as follows.

Definition 2.19. *The set of border elements $\mathcal{T}_H^B(\Omega) \subset \mathcal{T}_H(\Omega)$ is given by*

$$\mathcal{T}_H^B(\Omega) = \left\{ \tau \in \mathcal{T}_H(\Omega) \backslash \mathcal{T}_H^C(\Omega) \,\middle|\, \text{there exists } \tau' \in \mathcal{T}_H^C(\Omega) \text{ such that } \tau' \cap \tau \neq \emptyset \right\} \ . \tag{2.20}$$

See Figure 2-2 for an illustration of $\mathcal{T}_H^C(\Omega)$ and $\mathcal{T}_H^B(\Omega)$. So the border elements are the elements that are not themselves cut but share an edge or node with a cut element. The theory below will require several more assumptions about the coefficient function $\alpha(x)$ that are worth summarizing.

Assumption 2.20. *It is assumed that*

1. *the number of inclusions m is finite (Recall Definition 2.4).*
2. *the inclusions have C^∞ boundaries (required for regularity later in Theorem 2.22)*
3. *the finite element mesh $\mathcal{T}_H(\Omega)$ is fine enough such that there exists a $\tau \subset \Omega_i$ $i = 1, ..., m$ where $\tau \in \mathcal{T}_H(\Omega) \backslash \left(\mathcal{T}_H^C(\Omega) \cup \mathcal{T}_H^B(\Omega) \right)$.*

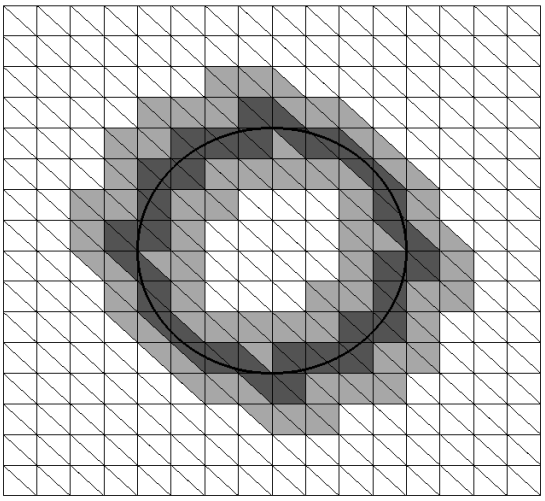


Figure 2-2: *An example mesh on a domain showing cut elements (dark grey), border elements (light grey) and then all remaining elements (white).*

4. *the finite element mesh is sufficiently fine such that given* $\mathcal{T}_1, \mathcal{T}_2 \subset \mathcal{T}_H^C(\Omega)$ *are any two sets of connected cut elements but* $\mathcal{T}_1$ *is not connected to* $\mathcal{T}_2$ *then*

$$2H < dist\,(\mathcal{T}_1, \mathcal{T}_2) = \min_{x \in \mathcal{T}_1,\ y \in \mathcal{T}_2} |x - y| \ .$$
$$2H < dist\,(\mathcal{T}_1, \partial\Omega)$$

The third assumption means that the finite element mesh $\mathcal{T}_H(\Omega)$ is fine enough to have at least one element sufficiently far from the boundary of each inclusion. The fourth assumption means the finite element mesh is sufficiently fine to have elements separating inclusions from each other and from the domain boundary. The fourth assuption also covers the case when the finite element mesh resolves part of the interface, when this occurs then the method of proof in the rest of this chapter requires that each subset of connected cut elements be surrounded by a closed curve of border elements and hence the need for each subset to be separated by at least two elements (a border element associated with each subset of cut elements).

2.2.2 A robust a priori error bound

The result we will prove in Theorem 2.58 shows that the error in the standard finite element approximation is robust with respect to the jumps in the coefficient $\alpha(x)$. This result is new. The elliptic interface problem has been studied by many people but the error estimates always have a constant that is dependent on the coefficient (for

example in [[77],Section 5], specifically equation (5.4), or [59]). This leads to an overly pessimistic estimate when the coefficient has large jumps. Also, the error estimate in Theorem 2.58 will show that with sufficient refinement of the mesh around the jumps in α we can restore an $O(H)$ convergence rate instead of $O(H^{\frac{1}{2}-\epsilon})$ in the energy norm even when the interface runs through the interior of mesh elements. The first tool we use is the optimality of the finite element solution in the energy norm.

Lemma 2.21 (Galerkin Optimality)**.**

$$|u - u_H|_{H^1(\Omega),\alpha} \leq |u - v_H|_{H^1(\Omega),\alpha} \qquad \textit{for all } v_H \in V_H \ . \tag{2.21}$$

Note that it is important that we have used the energy norm here since we obtain optimality independent of $\hat{\alpha}$. This is lost if we converted (2.21) into a statement about quasi-optimality in the norm $|\cdot|_{H^1(\Omega)}$.

In Lemma 2.21 we are free to construct any v_H in V_H on the right-hand side to obtain a good bound on the finite element error. Our choice of v_H (described below) will be obtained by interpolating u in a standard way on elements that are not cut by the interface and by interpolating u only on a subtriangle of the highest coefficient region on cut elements. After an averaging procedure to restore conformity we are able to obtain coefficient robustness. The proof requires the following regularity result.

Theorem 2.22 (Theorem B.1. of [27])**.** *Let Ω be either a smooth C^∞ bounded domain in $\mathbb{R}^2$ or a bounded convex polygon, let Ω contain inclusions Ω_i, i=1,2,...,m, each having a C^∞ boundary, and define $\Omega_0 = \Omega \backslash \cup_{i=1}^m \overline{\Omega}_m$ as in Definition 2.4. Consider Problem 2.2 and assume that either Case I (2.8) or Case II (2.9) holds. In addition, let $\Gamma = \bigcup_{i=1}^m \Gamma_i$ and let $\tilde{\Gamma}$ denote any closed C^∞ contour in Ω_0, which encloses all the Ω_i and let $\tilde{\Omega}_0$ be the domain with boundary $\Gamma \cup \tilde{\Gamma}$. Then we have*

$$|u|_{H^{s+2}(\Omega_i)} \lesssim \frac{1}{\alpha_i} \|f\|_{H^s(\Omega)}, \ \textit{for all } s \geq 0, \quad i = 1, 2, ..., m. \tag{2.22}$$

Moreover

$$|u|_{H^2(\Omega_0)} \lesssim \frac{1}{\alpha_0} \|f\|_{L_2(\Omega)}, \tag{2.23}$$

and

$$|u|_{H^{2+s}(\tilde{\Omega}_0)} \lesssim \frac{1}{\alpha_0} \|f\|_{H^s(\Omega)}, \ \textit{for all } s \geq 0. \tag{2.24}$$

The hidden constants depend on the distance of Γ from $\partial\Omega$.

The proof of this theorem is only given for a single inclusion in [27], we extend the

proof to multiple inclusions in Chapter 3 Theorem 3.23.

2.2.3 Approximation on cut elements

We will step through the construction of v_H to be inserted in the right hand side of (2.21) in stages. By Assumption 2.20 (4.) we have that each cut element can only contain regions in two inclusions. We designate the side of the interface with highest α coefficient in a cut element as Ω^- and then the other side as Ω^+. It is worth noting that this definition does not control the shape of the interface inside the element. An example of quite a variable interface is given in Figure 2-3. We clarify this with the following definition and assumption.

Definition 2.23. *Suppose $\tau \in \mathcal{T}_H^C(\Omega)$. Then by Assumption 2.20 (4.), τ can at most intersect two inclusions, Ω_i and Ω_j say. Let*

$$\Omega^- = \begin{cases} \Omega_i & \text{if } \alpha_i \geq \alpha_j \\ \Omega_j & \text{otherwise} \end{cases} \qquad \text{and} \qquad \Omega^+ = \begin{cases} \Omega_i & \text{if } \alpha_i < \alpha_j \\ \Omega_j & \text{otherwise} \end{cases}$$

then define

$$\tau^- := \tau \cap \Omega^- \qquad \text{and} \qquad \tau^+ := \tau \cap \Omega^+ \ . \tag{2.25}$$

Definition 2.24 (Definition 4.2.2 in Brenner and Scott [20])**.** *A domain γ is star shaped with respect to a ball B if, for all $x \in \gamma$, the closed convex hull of $\{x\} \cup B$ is a subset of γ.*

Assumption 2.25. *We assume that for each $\tau \in \mathcal{T}_H^C(\Omega)$, τ^- contains a triangle $K(\tau)$ of diameter $H_{K(\tau)}$ and τ^- is star shaped with respect to the largest inscribed ball in $K(\tau)$ of radius $\rho_{K(\tau)}$. We assume also that $K(\tau)$ is of comparable diameter to τ^- and $K(\tau)$ is shape regular, i.e. we assume*

$$H_{K(\tau)} \sim H_{\tau^-} \qquad \text{and} \qquad H_{K(\tau)}/\rho_{K(\tau)} \lesssim 1 \qquad \text{for all } \tau \in \mathcal{T}_H^C(\Omega) \ . \tag{2.26}$$

This still leaves freedom in how $K(\tau)$ is chosen and thus its size will enter into error estimates. An example $K(\tau)$ is shown in Figure 2-3. We will see later in this chapter that we will get a good error estimate when the area of $K(\tau)$ is comparable to the area of τ. This may not be possible though, so we introduce a parameter η_H that will appear in error estimates.

Definition 2.26. *Let*

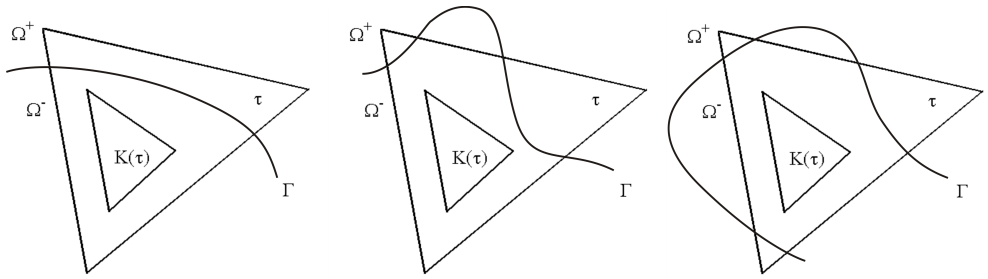


Figure 2-3: *Examples of a cut element τ showing the high coefficient side Ω^- and how the smaller shape regular element $K(\tau)$ can be defined.*

$$\eta_H = \max_{\tau \in \mathcal{T}_H^C(\Omega)} \frac{H_\tau}{H_{K(\tau)}}. \tag{2.27}$$

Following this definition we also need the following notation for the linear basis functions that span $V_H^{\mathbb{P}_1}$.

Definition 2.27. *For any triangle $\sigma \subset \mathbb{R}^2$ let x_i^σ for $i = 1, 2, 3$ be the nodes of σ, labelled in an anticlockwise fashion, and let ϕ_i^σ be the corresponding linear function such that*

$$\phi_i^\sigma(x_j^\sigma) = \delta_{ij}.$$

when restricted to σ, ϕ_i^σ is the standard nodal basis function on σ. However ϕ_i^σ is defined on the whole of $\mathbb{R}^2$, not just σ. This is trivial but important for the argument in the rest of this chapter.

With the introduction of basis functions that span the approximation space we also note an important feature of modern finite element methods. The basis functions are usually chosen to form a partition of unity.

Definition 2.28. *A set of functions $\{\Phi_i\}_{i=1}^m$ over the space Ω forms a partition of unity if at any point $x \in \Omega$:*

1. *Finitely many functions are non-zero and,*
2. *all the functions sum to 1.*

The partition of unity property is important because it allows results (such as integration) calculated in a local area to be extended to the whole domain.

Definition 2.29. *For a triangle σ define the nodal interpolant on the plane as*

$$I_\sigma u(x) = \sum_{i=1}^{3} u(x_i^\sigma)\phi_i^\sigma(x) \qquad x \in \mathbb{R}^2 \ , \tag{2.28}$$

where x_i^σ are the nodes and ϕ_i^σ are the linear basis functions in Definition 2.27.

We will also require some tools that show that shape regularity is preserved under an affine transformation based on a shape regular triangle (see Appendix A). These tools and the analysis in the remaining part of this chapter first requires the definition of an affine map for an element of the mesh.

Definition 2.30. *Let $\hat{\tau}$ denote the unit triangle*

$$\hat{\tau} = \left\{ x \in \mathbb{R}^2 \,|\, x_1 \geq 0,\ x_2 \geq 0 \text{ and } x_1 + x_2 \leq 1 \right\} \ .$$

For any triangle $\sigma \subset \mathbb{R}^2$ let F_σ be the affine map which maps $\hat{\tau}$ to σ, i.e.

$$F_\sigma(\hat{x}) = A_\sigma \hat{x} + b_\sigma \tag{2.29}$$

for any $\hat{x} \in \hat{\tau}$ where

$$A_\sigma = \begin{bmatrix} x_2^\sigma - x_1^\sigma & x_3^\sigma - x_1^\sigma \end{bmatrix} \qquad \text{and} \qquad b_\sigma = x_1^\sigma \ . \tag{2.30}$$

Lemma 2.31. *For any triangle σ*

$$\left|A_\sigma^T\right|_2 = |A_\sigma|_2 \lesssim H_\sigma \qquad \text{and} \qquad \left|A_\sigma^{-T}\right|_2 = \left|A_\sigma^{-1}\right|_2 \lesssim \rho_\sigma^{-1} \tag{2.31}$$

where $|\cdot|_2$ denotes the matrix norm induced by the Euclidean norm $|\cdot|_2$ on vectors.

Proof. This is a classical result where the upper bound on the matrix 2-norm can be found in Theorem 3.1.3 of Ciarlet [29] and the equivalence of the transpose matrix 2-norm can be found in Golub and Van Loan [41]. □

Now we use the estimates of Lemma 2.31 to bound the linear functions ϕ_i^σ and their gradients.

Lemma 2.32. *Suppose $\gamma \subset \mathbb{R}^2$ is a domain and $\sigma \subset \gamma$ is a triangle then we have the bounds*

$$\|\nabla \phi_i^\sigma\|_{L_\infty(\gamma)} \lesssim \frac{1}{\rho_\sigma} \qquad \text{and} \qquad \|\phi_i^\sigma\|_{L_\infty(\gamma)} \lesssim \frac{H_\gamma}{\rho_\sigma} \ . \tag{2.32}$$

Proof. We denote the nodes on the unit triangle $\hat{\tau}$ by $x_i^{\hat{\tau}}$ $i = 1, 2, 3$ and assume F_σ maps $x_i^{\hat{\tau}}$ to x_i^σ. Then we have

$$\phi_i^{\hat{\tau}}(\hat{x}) = \phi_i^{\sigma}\left(F_\sigma(\hat{x})\right)$$

and taking the gradient gives

$$\nabla_{\hat{\tau}}\phi_i^{\hat{\tau}}(\hat{x}) = A_\sigma^T\left(\nabla_\sigma\phi_i^{\sigma}\left(F_\sigma(\hat{x})\right)\right)$$

where $\nabla_{(\cdot)}$ is the gradient in the corresponding coordinates. Since ϕ_i^σ is linear then its gradient is constant and so the above equation is independent of $\hat{x}$. Therefore

$$|\nabla_\sigma\phi_i^\sigma|_2 = \left|A_\sigma^{-T}\nabla_{\hat{\tau}}\phi_i^{\hat{\tau}}\right|_2 \leq \left|A_\sigma^{-T}\right|_2\left|\nabla_{\hat{\tau}}\phi_i^{\hat{\tau}}\right|_2 \lesssim \rho_\sigma^{-1}$$

by Lemma 2.31 and also using the fact that $\left|\nabla_{\hat{\tau}}\phi_i^{\hat{\tau}}\right|_2 \lesssim 1$ because $\phi_i^{\hat{\tau}}$ is a basis function on the unit triangle $\hat{\tau}$. Now for any $i \neq j$ and any $x \in \gamma$

$$\phi_i^\sigma(x) = \phi_i^\sigma(x) - \phi_i^\sigma(x_j^\sigma) = \left(\nabla_\sigma\phi_i^\sigma\right)^T\left(x - x_j^\sigma\right)$$

by Taylor's theorem. Since $x, x_j^\sigma \in \gamma$, this implies that

$$\|\phi_i^\sigma\|_{L_\infty(\gamma)} \lesssim \frac{H_\gamma}{\rho_\sigma} \ .$$

□

The key idea to obtaining a robust finite element error is to interpolate on the high contrast side τ^- (recall (2.25)). To this end we employ $I_{K(\tau)}u$ as defined in Definition 2.29 and recall that it is defined on all of $\mathbb{R}^2$ (and hence on all of τ). Now to bound $u - I_{K(\tau)}u$ in the energy norm on τ we will need the following proposition.

Proposition 2.33. *Let $\tau \in \mathcal{T}_H^C(\Omega)$ and γ be a domain such that $K(\tau) \subset \overline{\gamma} \subseteq \tau$. Then*

$$(i) \quad \left\|I_{K(\tau)}v\right\|_{L_\infty(\gamma)} \lesssim \frac{H_\gamma}{\rho_{K(\tau)}}\|v\|_{L_\infty(K(\tau))} \tag{2.33}$$

$$(ii) \quad \left|I_{K(\tau)}v\right|_{H^1(\gamma)} \lesssim \frac{H_\gamma}{\rho_{K(\tau)}}\|v\|_{L_\infty(K(\tau))} \tag{2.34}$$

for all $v \in C(\tau)$.

Proof. Firstly for brevity let $K := K(\tau)$ and note that

$$I_K v = \sum_{j=1}^{3} v(x_j^K)\phi_j^K \ .$$

This implies that

$$\|I_K v\|_{L_\infty(\gamma)} \lesssim \|v\|_{L_\infty(K)} \max_{j=1,2,3} \left\|\phi_j^K\right\|_{L_\infty(\gamma)} \lesssim \frac{H_\gamma}{\rho_K} \|v\|_{L_\infty(K)} \ ,$$

by Lemma 2.32. Also

$$|I_K v|_{H^1(\gamma)} \lesssim \|v\|_{L_\infty(K)} \max_{j=1,2,3} \left|\phi_j^K\right|_{H^1(\gamma)} \lesssim \frac{H_\gamma}{\rho_K} \|v\|_{L_\infty(K)} \ ,$$

again using Lemma 2.32. □

This now gives us enough tools to obtain a robust estimate for $\left|u - I_{K(\tau)}u\right|_{H^1(\tau),\alpha}$. We will proceed by writing

$$\left|u - I_{K(\tau)}u\right|^2_{H^1(\tau),\alpha} = \alpha^- \left|u - I_{K(\tau)}u\right|^2_{H^1(\tau^-)} + \alpha^+ \left|u - I_{K(\tau)}u\right|^2_{H^1(\tau^+)} . \tag{2.35}$$

where $\alpha^- = \alpha|_{\tau^-}$ and $\alpha^+ = \alpha|_{\tau^+}$ and estimating each term on the right-hand side separately. First we obtain the bound on the high coefficient side, τ^-. To do this we will make use of two common results in finite elements.

Theorem 2.34 (Theorem 3.1.2 of Ciarlet [29])**.** *Let $\widehat{\Omega} \subset \mathbb{R}^2$ and $F(\hat{x}) = A\hat{x} + b$ be an affine map such that $\Omega = F(\widehat{\Omega})$ (meaning for any $\hat{x} \in \widehat{\Omega}$ then $F(\hat{x}) \in \Omega$). If $v \in H^m(\Omega)$ for some interger $m \geq 0$, then $\hat{v} = v \cdot F \in H^m(\widehat{\Omega})$ and there exists a constant $C = C(m)$ such that*

$$|\hat{v}|_{H^m(\widehat{\Omega})} \leq C \, |A|_2^m \, |\det A|^{-\frac{1}{2}} \, |v|_{H^m(\Omega)} \ . \tag{2.36}$$

Analogously, one has

$$|v|_{H^m(\Omega)} \leq C \left|A^{-1}\right|_2^m |\det A|^{\frac{1}{2}} \, |\hat{v}|_{H^m(\widehat{\Omega})} \ . \tag{2.37}$$

We apply Lemma 2.31 to Theorem 2.34 to get the following corollary.

Corollary 2.35. *For a triangle σ and resulting affine map F_σ (see Definition 2.30) we have, for all $v \in H^m(\Omega)$*

$$|\hat{v}|_{H^m(\widehat{\Omega})} \lesssim H_\sigma^m \, |\det A_\sigma|^{-\frac{1}{2}} \, |v|_{H^m(\Omega)} \tag{2.38}$$

and

$$|v|_{H^m(\Omega)} \lesssim \rho_\sigma^{-m} \, |\det A_\sigma|^{\frac{1}{2}} \, |\hat{v}|_{H^m(\widehat{\Omega})} \tag{2.39}$$

Proof. These result from using Lemma 2.31 by substituting $|A_\sigma|_2 \lesssim H_\sigma$ into (2.36) and $\left|A_\sigma^{-1}\right|_2 \lesssim \rho_\sigma^{-1}$ into (2.37). □

The other common finite element tool that we use is the Bramble-Hilbert Lemma.

Lemma 2.36 (Lemma 4.3.8 in Brenner and Scott [20])**.** *Let $\sigma \subset \mathbb{R}^2$ be a domain and let B be a ball of radius ρ in σ, such that σ is star-shaped with respect to B and such that $\rho > \frac{1}{2}\rho_\sigma$. Then there exists a polynomial q of order $m-1$ such that for each $u \in H^m(\sigma)$ we have the bound*

$$|u-q|_{H^k(\sigma)} \lesssim H_\sigma^{m-k} |u|_{H^m(\sigma)} \qquad \text{for } k = 0, 1, ..., m \ , \tag{2.40}$$

with the hidden constant independent of u.

We also use the following lemma based on rescaling a Sobolev embedding result.

Lemma 2.37. *Let σ be a shape regular domain (i.e. $\rho_\sigma \sim H_\sigma$). Then for any $v \in L_\infty(\sigma)$*

$$\|v\|_{L_\infty(\sigma)} \lesssim H_\sigma^{-1} \|v\|_{L_2(\sigma)} + |v|_{H^1(\sigma)} + H_\sigma^\epsilon |v|_{H^{1+\epsilon}(\sigma)} \tag{2.41}$$

for arbitrary $\epsilon > 0$, and hidden constant independent of ϵ.

Proof. Firstly let $\hat{\sigma} = \frac{1}{H_\sigma}\sigma$ and thus $H_{\hat{\sigma}} \sim 1$ by the shape regularity of σ. Then let $\hat{v}(x) = v(\frac{x}{H_\sigma})$ so

$$\|v\|_{L_\infty(\sigma)} = \|\hat{v}\|_{L_\infty(\hat{\sigma})} \lesssim \|\hat{v}\|_{H^{1+\epsilon}(\hat{\sigma})}$$

by the Sobolev embedding theorem for $\epsilon > 0$. Note that the hidden constant depends on the size of $\hat{\sigma}$ but this is $O(1)$. By a change of variables with $\hat{x} = \frac{x}{H_\sigma}$ we obtain

$$\|\hat{v}\|_{L_2(\hat{\sigma})} = H_\sigma^{-1} \|v\|_{L_2(\sigma)} \ , \quad |\hat{v}|_{H^1(\hat{\sigma})} = |v|_{H^1(\sigma)} \ , \quad |\hat{v}|_{H^{1+\epsilon}(\hat{\sigma})} = H_\sigma^\epsilon |v|_{H^{1+\epsilon}(\sigma)}$$

where $H^{1+\epsilon}$ is equipped with the Slobodeckiĭ seminorm [p74 McLean [68]]. □

Before this point we have used the notation $\widehat{\cdot}$ to denote the pullback under a general affine mapping, from now on we will restrict to using this notation for an affine map associated with an element $\tau \in \mathcal{T}_H(\Omega)$.

Lemma 2.38. *For any $\tau \in \mathcal{T}_H^C(\Omega)$ with corresponding $K(\tau)$ and for any $v \in C(K(\tau))$ we have*

$$\left(\widehat{I_{K(\tau)} v}\right)(\hat{x}) = \left(I_{\widehat{K(\tau)}} \widehat{v}\right)(\widehat{x})$$

where $\widehat{\cdot}$ denotes the pullback under F_τ.

Proof. Note that

$$\begin{aligned}\left(I_{K(\tau)}v\right)\left(F_\tau(\widehat{x})\right) &= \sum_{j=1}^{3} v(x_j^{K(\tau)})\phi_j^{K(\tau)}\left(F_\tau(\widehat{x})\right) \\ &= \sum_{j=1}^{3} v(F_\tau(\hat{x}_j^{\widehat{K(\tau)}}))\phi_j^{K(\tau)}\left(F_\tau(\widehat{x})\right) \\ &= \sum_{j=1}^{3} \hat{v}(\hat{x}_j^{\widehat{K(\tau)}})\phi_j^{\widehat{K(\tau)}}\left(\widehat{x}\right) \ .\end{aligned}$$

□

We show that approximating by piecewise linears on $K(\tau)$ and extending the approximation to all of τ^- gives an optimal error estimate on τ^-. Using the previous three lemmas we can prove the following error estimate on τ^-.

Theorem 2.39. *For $\tau \in \mathcal{T}_H^C(\Omega)$, under Assumptions 2.15, 2.20 and 2.25 we have that*

$$\left|u - I_{K(\tau)}u\right|_{H^1(\tau^-)} \lesssim H_\tau \left|u\right|_{H^2(\tau^-)} \ . \tag{2.42}$$

Proof. By Lemma 2.38

$$\left(u - I_{K(\tau)}u\right)\hat{}\,(\hat{x}) = \left(\hat{u} - I_{\widehat{K(\tau)}}\hat{u}\right)(\hat{x}) = (\hat{u} - \hat{q}) - I_{\widehat{K(\tau)}}(\hat{u} - \hat{q})$$

for all $\hat{q} \in \mathbb{P}_1$. Hence by Proposition 2.33 using $\gamma = \widehat{\tau^-}$ and $\widehat{K(\tau)} \subset \widehat{\tau^-} \subset \widehat{\tau}$ we have

$$\begin{aligned}\left|\left(u - I_{K(\tau)}u\right)\hat{}\,\right|_{H^1(\widehat{\tau^-})} &\leq |\hat{u} - \hat{q}|_{H^1(\widehat{\tau^-})} + \left|I_{\widehat{K(\tau)}}(\hat{u} - \hat{q})\right|_{H^1(\widehat{\tau^-})} \\ &\leq |\hat{u} - \hat{q}|_{H^1(\widehat{\tau^-})} + \frac{H_{\widehat{\tau^-}}}{\rho_{\widehat{K(\tau)}}} \|\hat{u} - \hat{q}\|_{L_\infty(\widehat{K(\tau)})}\end{aligned}$$

Note that by Lemma A.1 (using $\gamma = \tau^-$ and $\sigma = \tau$) and A.2 (using $\gamma = K(\tau)$ and $\sigma = \tau$)

$$\frac{H_{\widehat{\tau^-}}}{\rho_{\widehat{K(\tau)}}} \lesssim \frac{H_{\tau^-}}{\rho_\tau} \cdot \frac{H_\tau}{\rho_{K(\tau)}}$$

Hence, by Assumption 2.15 and 2.25,

$$\frac{H_{\widehat{\tau^-}}}{\rho_{\widehat{K(\tau)}}} \lesssim \frac{H_{\tau^-}}{\rho_{K(\tau)}} \lesssim \frac{H_{\tau^-}}{H_{K(\tau)}} \cdot \frac{H_{K(\tau)}}{\rho_{K(\tau)}} \lesssim 1 \ .$$

Therefore,

$$\begin{aligned}\left|\left(u - I_{K(\tau)}u\right)\hat{}\,\right|_{H^1(\widehat{\tau^-})} &\lesssim |\hat{u}-\hat{q}|_{H^1(\widehat{\tau^-})} + \|\hat{u}-\hat{q}\|_{L_\infty(\widehat{K(\tau)})} \\ &\leq |\hat{u}-\hat{q}|_{H^1(\widehat{\tau^-})} + \|\hat{u}-\hat{q}\|_{L_\infty(\widehat{\tau^-})} \\ &\lesssim H_{\widehat{\tau^-}}^{-1}\,\|\hat{u}-\hat{q}\|_{L_2(\widehat{\tau^-})} + |\hat{u}-\hat{q}|_{H^1(\widehat{\tau^-})} + H_{\widehat{\tau^-}}^{\epsilon}\,|\hat{u}-\hat{q}|_{H^{1+\epsilon}(\widehat{\tau^-})}\end{aligned}$$

using Lemma 2.37 since τ^- is shape regular. Now by Lemma 2.36 we have

$$\begin{aligned}&(i) \quad H_{\widehat{\tau^-}}^{-1}\,\|\hat{u}-\hat{q}\|_{L_2(\widehat{\tau^-})} \lesssim H_{\widehat{\tau^-}}\,|\hat{u}|_{H^2(\widehat{\tau^-})} \lesssim |\hat{u}|_{H^2(\widehat{\tau^-})} \\ &(ii) \quad |\hat{u}-\hat{q}|_{H^1(\widehat{\tau^-})} \lesssim H_{\widehat{\tau^-}}\,|\hat{u}|_{H^2(\widehat{\tau^-})} \lesssim |\hat{u}|_{H^2(\widehat{\tau^-})} \\ &(iii) \quad H_{\widehat{\tau^-}}^{\epsilon}\,|\hat{u}-\hat{q}|_{H^{1+\epsilon}(\widehat{\tau^-})} \lesssim |\hat{u}|_{H^2(\widehat{\tau^-})}\end{aligned}$$

since $H_{\widehat{\tau^-}} \leq H_{\widehat{\tau}} \lesssim 1$. Therefore

$$\left|\left(u - I_{K(\tau)}u\right)\hat{}\,\right|_{H^1(\widehat{\tau^-})} \lesssim |\hat{u}|_{H^2(\widehat{\tau^-})}\ .$$

Combining this with (2.38) and (2.39) we obtain

$$\begin{aligned}\left|u - I_{K(\tau)}u\right|_{H^1(\tau^-)} &\lesssim \rho_\tau^{-1}\,|\det\,A_\tau|^{\frac{1}{2}}\left|\left(u - I_{K(\tau)}u\right)\hat{}\,\right|_{H^1(\widehat{\tau^-})} \\ &\lesssim \rho_\tau^{-1}\,|\det\,A_\tau|^{\frac{1}{2}}\,|\hat{u}|_{H^2(\widehat{\tau^-})} \\ &\lesssim \rho_\tau^{-1}\,|\det\,A_\tau|^{\frac{1}{2}}\,H_\tau^2\,|\det\,A_\tau|^{-\frac{1}{2}}\,|u|_{H^2(\tau^-)} \\ &\lesssim H_\tau\,|u|_{H^2(\tau^-)}\ ,\end{aligned}$$

using the shape regularity of τ. □

Since $u \notin H^2$ on all of τ and since $I_{K(\tau)}u$ samples u only on τ^- we cannot use the pullback to obtain the analogue of (2.42) on τ^+. Instead we use an approximation result by Scott and Zhang in [77] to obtain a lower order result. This requires the definition of a quasi-interpolant.

Definition 2.40. *Given a function $v \in H^1(\tau)$, the conventional nodal polynomial interpolant $Iv \in \mathbb{P}_p$ is a polynomial function of order p where $Iv(x_i) = v(x_i)$ for all the nodes $x_i \in \tau$ (see Definition 2.29). Denote the diameter of τ as H, then*

$$\|v - Iv\|_{H^1(\tau)} \to 0 \qquad \text{as } H \to 0.$$

A quasi-interpolant has the same properties but does not necessarily interpolate the nodes, i.e. it may be the case that $Iv(x_i) \neq v(x_i)$ for any of the nodes $x_i \in \tau$.

Theorem 2.41 (Section 4 of [77]). *For any $\tau \in \mathcal{T}_H(\Omega)$ and $v \in H^{\frac{3}{2}-\epsilon}(\Omega)$ there exists a quasi-interpolant $\Pi v \in V_H^{\mathbb{P}_1}$ such that*

$$\|v - \Pi v\|_{H^k(\tau)} \lesssim H_\tau^{m-k} |v|_{H^m(S_\tau)} \tag{2.43}$$

for $0 \leq k \leq m \leq \frac{3}{2} - \epsilon$ and where S_τ is the set of neighbours to τ given by

$$S_\tau = \left\{ \tau' \in \mathcal{T}_H(\Omega) \mid \tau' \cap \tau \neq \emptyset \right\} . \tag{2.44}$$

This theorem uses a version of the Bramble-Hilbert lemma developed in [35] for fractional order Sobolev spaces. Using this we can proceed in a similar way to Theorem 2.39 but without using the pullback to the unit triangle, to obtain the following lower order estimate.

Theorem 2.42. *For $\tau \in \mathcal{T}_H(\Omega)$, under Assumptions 2.20, 2.15 and 2.25 we have that*

$$\left|u - I_{K(\tau)}u\right|_{H^1(\tau^+)} \lesssim H_\tau^{\frac{1}{2}-\epsilon} (1 + \eta_H) |u|_{H^{\frac{3}{2}-\epsilon}(S_\tau)} . \tag{2.45}$$

Proof. By Theorem 2.41 there exists a $q \in \mathbb{P}_1$ such that for all $0 \leq k \leq \frac{3}{2} - \epsilon$

$$\|u - q\|_{H^k(\tau)} \lesssim H_\tau^{\frac{3}{2}-\epsilon-k} |u|_{H^{\frac{3}{2}-\epsilon}(S_\tau)} . \tag{2.46}$$

Since $I_{K(\tau)}q = q$ for $q \in \mathbb{P}_1$ we have

$$\begin{aligned}
\left|u - I_{K(\tau)}u\right|_{H^1(\tau^+)} &= \left|(u-q) - I_{K(\tau)}(u-q)\right|_{H^1(\tau^+)} \\
&\leq |u-q|_{H^1(\tau^+)} + \left|I_{K(\tau)}(u-q)\right|_{H^1(\tau^+)} \\
&\leq |u-q|_{H^1(\tau)} + \left|I_{K(\tau)}(u-q)\right|_{H^1(\tau)} \\
&\lesssim |u-q|_{H^1(\tau)} + \frac{H_\tau}{\rho_{K(\tau)}} \|u-q\|_{L_\infty(K(\tau))}
\end{aligned}$$

by Proposition 2.33 (ii). Hence using Definition 2.26 and Assumption 2.25,

$$\begin{aligned}
\left|u - I_{K(\tau)}u\right|_{H^1(\tau^+)} &\lesssim (1+\eta_H) \left(|u-q|_{H^1(\tau)} + \|u-q\|_{L_\infty(\tau)} \right) \\
&\lesssim (1+\eta_H) \left(H_\tau^{-1} \|u-q\|_{L_2(\tau)} + |u-q|_{H^1(\tau)} + H_\tau^\epsilon |u-q|_{H^{1+\epsilon}(\tau)} \right) ,
\end{aligned}$$

where we have used Lemma 2.37. By (2.46) we then have

$$\begin{aligned}\left|u - I_{K(\tau)}u\right|_{H^1(\tau^+)} &\lesssim (1+\eta_H)\left(H_\tau^{-1+\frac{3}{2}-\epsilon} + H_\tau^{\frac{1}{2}-\epsilon} + H_\tau^{\epsilon+\frac{1}{2}-2\epsilon}\right)|u|_{H^{\frac{3}{2}-\epsilon}(S_\tau)} \\ &\lesssim H_\tau^{\frac{1}{2}-\epsilon}(1+\eta_H)\,|u|_{H^{\frac{3}{2}-\epsilon}(S_\tau)}\end{aligned}$$

as required. □

Combining Theorems 2.39 and 2.42 we get the following corollary.

Corollary 2.43. *For $\tau \in \mathcal{T}_H^C(\Omega)$,*

$$\left|u - I_{K(\tau)}u\right|_{H^1(\tau),\alpha} \lesssim H_\tau^{\frac{1}{2}-\epsilon}\left(\alpha|_{\tau^-}\, H_\tau^{1+2\epsilon}\,|u|^2_{H^2(\tau^-)} + (1+\eta_H)^2\,|u|^2_{H^{\frac{3}{2}-\epsilon}(S_\tau)}\right)^{\frac{1}{2}}. \tag{2.47}$$

Proof. Substitute equations (2.42) and (2.45) into (2.35) and take the square root of both sides. Note also that $\alpha|_{\tau^+} \lesssim 1$ (recall Case I (2.8) and Case II (2.9)). □

2.3 A priori error bound for cut and border elements

We now need to construct a candidate for v_H in (2.21) to obtain a robust upper bound on $|u - u_H|_{H^1(\Omega),\alpha}$. Based on what we saw in the previous subsections, we may be inclined to define v_H on cut elements as $v_H|_\tau = I_{K(\tau)}u$. However such a v_H will not necessarily be continuous across element edges. Consider two elements τ_1 and τ_2 that share an edge e with nodes x_1, x_2 (see Figure 2-4) and consider $K(\tau_i)$ as in Figure 2-4. Since $K(\tau_1)$ and $K(\tau_2)$ have a common edge along e, $I_{K(\tau_1)}u$ and $I_{K(\tau_2)}u$ are equal along $K(\tau_1) \cap K(\tau_2)$ and hence equal on all of e. However $I_{K(\tau_3)}u$ is not necessarily equal to $I_{K(\tau_2)}u$ because $K(\tau_2)$ and $K(\tau_3)$ only share a node.

In the rest of this section we provide a technical solution to this problem. We utilise the ideas developed so far of creating interpolants in the high coefficient areas but then glue them together using a modification to create a continuous v_H.

Definition 2.44. *For $x_j \in \mathcal{N}(\mathcal{T}_H(\Omega))$ we define*

$$S^C(x_j) = \left\{\tau \in \mathcal{T}_H^C(\Omega) \,|\, x_j \in \tau\right\}. \tag{2.48}$$

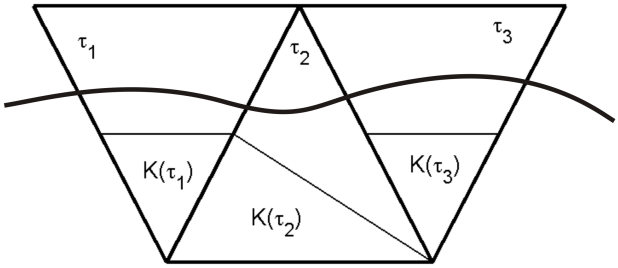


Figure 2-4: *An illustration of why $I_{K(\tau)}u$ cannot be used to define a globally continuous function. Here $I_{K(\tau_1)}u$ and $I_{K(\tau_2)}u$ are continuous across the edge $\tau_1 \cap \tau_2$, but $I_{K(\tau_2)}u$ and $I_{K(\tau_2)}u$ are not continuous across the edge $\tau_2 \cap \tau_3$.*

Then for $v \in C(\overline{\Omega})$, define

$$\beta_j(v) = \begin{cases} \frac{1}{M_j} \sum\limits_{\tau \in S^C(x_j)} \left(I_{K(\tau)}v\right)(x_j) & \text{if } S^C(x_j) \neq \emptyset \quad \text{where} \quad M_j = \left|S^C(x_j)\right| \\ v(x_j) & \text{otherwise} \quad . \end{cases} \tag{2.49}$$

We then define a function in $V_H^{\mathbb{P}_1}$ using these nodal weights. For this we need some notation.

Notation 2.45. *Let $\mathcal{G}_\tau$ denote the local to global mapping that maps the local indices $\{1,2,3\}$ to the global indices $\{i,j,k\}$ where $x_i, x_j, x_k \in \mathcal{N}(\mathcal{T}_H(\Omega))$ are the nodes of τ.*

Definition 2.46. *Define the operator J_H by*

$$J_H u|_\tau(x) = \sum_{i=1}^{3} \beta_{\mathcal{G}_\tau(i)}(u)\phi_i^\tau(x) \ , \tag{2.50}$$

for all $\tau \in \mathcal{T}_H(\Omega)$ and $v \in C(\overline{\Omega})$.

In other words, $J_H v$ is the continuous function that interpolates the values $\beta_j(v)$ for all $x_j \in \mathcal{N}(\mathcal{T}_H(\Omega))$ and is linear on each $\tau \in \mathcal{T}_H(\Omega)$. The idea of the nodal weights $\beta_j(u)$ is to average the values of $I_{K(\tau)}u$ over all cut neighbours at each node of a cut element, e.g. if $x \in \mathcal{N}(\mathcal{T}_H(\Omega))$ is a node of two cut elements τ_1 and τ_2 then $J_H u(x) = \frac{1}{2}\left(I_{K(\tau_1)}u(x) + I_{K(\tau_2)}u(x)\right)$. From the definition of J_H we obtain the following lemma.

Lemma 2.47.

$$J_H p = p \qquad \text{for all} \quad p \in V_H^{\mathbb{P}_1} \ . \tag{2.51}$$

Proof. Let $p \in V_H^{\mathbb{P}_1}$ and $x_j \in \mathcal{N}(\mathcal{T}_H(\Omega))$. If $S^C(x_j) \neq \emptyset$, consider any $\tau \in S^C(x_j)$. Clearly $I_{K(\tau)}p = p$ and so $\beta_j(p) = p(x_j)$. Also if $S^C(x_j) = \emptyset$ then $\beta_j(p) = p(x_j)$. Hence

$$J_H p|_\tau = \sum_{i=1}^{3} p(x_{\mathcal{G}_\tau(i)})\phi_i^\tau = p|_\tau \ .$$

□

Now we can proceed to prove the global analogue of Corollary 2.43. This will be given in Theorem 2.50 below but first we need the following two lemmas.

Lemma 2.48. *For all $x_j \in \mathcal{N}(\mathcal{T}_H(\Omega))$ and $v \in C(\overline{\Omega})$*

$$|\beta_j(v)| \lesssim \begin{cases} \eta_H \|v\|_{L_\infty(S^C(x_j)\cap\Omega^-)} & \text{if } S^C(x_j) \neq \emptyset \\ \|v\|_{L_\infty(\tau)} & \text{otherwise} \end{cases} , \tag{2.52}$$

for any $\tau \in \mathcal{T}_H(\Omega)$ such that $x_j \in \tau$.

Proof. First suppose that $S^C(x_j) \neq \emptyset$ then

$$\begin{aligned} |\beta_j(v)| \leq \frac{1}{M_j} \sum_{\tau\in S^C(x_j)} \left|\left(I_{K(\tau)}v\right)(x_j)\right| &\leq \frac{1}{M_j} \sum_{\tau\in S^C(x_j)} \left\|\left(I_{K(\tau)}v\right)\right\|_{L_\infty(\tau)} \\ &\lesssim \frac{1}{M_j} \sum_{\tau\in S^C(x_j)} \eta_H \|v\|_{L_\infty(K(\tau))} \ , \end{aligned}$$

by Proposition 2.33 and recalling Definition 2.26 and Assumption 2.25. Consequently, also using the fact that for $\tau \in S^C(x_j)$ $K(\tau) \subset \tau^- \subset S^C(x_j) \cap \Omega^-$.

$$|\beta_j(v)| \lesssim \eta_H \|v\|_{L_\infty(S^C(x_j)\cap\Omega^-)} \ .$$

On the other hand if $S^C(x_j) = \emptyset$ then trivially $|\beta_j(v)| = |v(x_j)| \leq \|v\|_{L_\infty(\tau)}$ for any $\tau \in \mathcal{T}_H(\Omega)$ that contains x_j. □

Lemma 2.49. *Let $\widehat{\cdot}$ denote the pullback under the affine map $F_\tau(\hat{x}) = A_\tau\hat{x} + b_\tau$. Then*

$$\widehat{J_H v}\Big|_{\widehat{\tau}}(\hat{x}) = \sum_{i=1}^{3} \beta_{\mathcal{G}_\tau(i)}(v)\phi_i^{\widehat{\tau}}(\hat{x}) \tag{2.53}$$

for any $v \in C(\overline{\Omega})$ and $\tau \in \mathcal{T}_H(\Omega)$. Also for any domain γ such that $\gamma \subset \tau$ we have

$$\left|\widehat{J_H v}\right|_{H^1(\widehat{\gamma})} \lesssim \frac{H_{\widehat{\gamma}}}{\rho_{\widehat{\tau}}} \max_{i=1,2,3} \left|\beta_{\mathcal{G}_\tau(i)}(v)\right| \ . \tag{2.54}$$

Proof. Using the definition of the pullback of ϕ_i^τ to $\phi_i^{\hat{\tau}}$ we obtain

$$\left|\widehat{J_H v}\right|_{H^1(\widehat{\gamma})} \leq \max_{i=1,2,3} \left|\beta_{\mathcal{G}_\tau(i)}(v)\right| \left|\phi_i^{\widehat{\tau}}\right|_{H^1(\widehat{\gamma})} \lesssim \frac{H_{\widehat{\gamma}}}{\rho_{\widehat{\tau}}} \max_{i=1,2,3} \left|\beta_{\mathcal{G}_\tau(i)}(v)\right|$$

by Lemma 2.32. □

In the same way as (2.35) we decompose the energy norm error between u and $J_H u$ on the high and low coefficient sides of an interface running through a cut element,

$$|u - J_H u|^2_{H^1(\tau),\alpha} = \alpha^- |u - J_H u|^2_{H^1(\tau^-)} + \alpha^+ |u - J_H u|^2_{H^1(\tau^+)} \, . \tag{2.55}$$

We also split the border elements around each interface into two groups. Given a particular interface Γ_i we have a subset of cut elements that intersect that interface (see Figure 2-2). By Definition 2.23, each cut element is split into τ^- and τ^+ for the high and low coefficient sides respectively. Similarly we can then also split the border elements into those on the high and low coefficient sides as:

$$\begin{aligned}
\mathcal{T}_H^B(\Omega^-) &:= \left\{\tau' \in \mathcal{T}_H^B(\Omega) \,|\, \text{there exists a } \tau \in \mathcal{T}_H^C(\Omega) \text{ such that } \tau' \cap \tau^- \neq \emptyset\right\} \ , \\
\mathcal{T}_H^B(\Omega^+) &:= \left\{\tau' \in \mathcal{T}_H^B(\Omega) \,|\, \text{there exists a } \tau \in \mathcal{T}_H^C(\Omega) \text{ such that } \tau' \cap \tau^+ \neq \emptyset\right\} \ .
\end{aligned}$$

Then we proceed with a modification of the argument in Theorem 2.39 to bound the error in the high coefficient region τ^- of (2.55).

Theorem 2.50. *Suppose $\tau \in \mathcal{T}_H^C(\Omega)$ and define $\gamma := \tau^-$. Alternatively, if $\tau \in \mathcal{T}_H^B(\Omega^-)$, then define $\gamma := \tau$. Then*

$$|u - J_H u|_{H^1(\gamma)} \lesssim H_\tau \left(1 + \eta_H\right) |u|_{H^2(S_\tau \cap \Omega^-)} \tag{2.56}$$

with S_τ as defined in (2.44).

Proof. First, by using Lemma 2.47 we have

$$(u - J_H u)\hat{} \ = \widehat{u} - \widehat{J_H u} = (\widehat{u} - \widehat{q}) - (J_H(u - q))\hat{} \ ,$$

for $q \in \mathbb{P}_1$. Lemma 2.49 also gives

$$|(u - J_H u)\hat{} \,|_{H^1(\widehat{\gamma})} \leq |\widehat{u} - \widehat{q}|_{H^1(\widehat{\gamma})} + \max_{i=1,2,3} \left|\beta_{\mathcal{G}_\tau(i)}(u - q)\right|$$

since $H_{\widehat{\gamma}}/\rho_{\widehat{\tau}} \leq H_{\widehat{\tau}}/\rho_{\widehat{\tau}} \lesssim 1$. Now $x_{\mathcal{G}_\tau(i)}$ is a node of τ, and $S^C(x_{\mathcal{G}_\tau(i)}) \subset S_\tau$. So by Lemma 2.48 we obtain

$$\begin{aligned}|(u - J_H u)^\wedge|_{H^1(\widehat{\gamma})} &\lesssim |\widehat{u} - \widehat{q}|_{H^1(\widehat{\gamma})} + \eta_H \|u - q\|_{L_\infty(S_\tau \cap \Omega^-)} \\ &= |\widehat{u} - \widehat{q}|_{H^1(\widehat{\gamma})} + \eta_H \|\widehat{u} - \widehat{q}\|_{L_\infty(\widehat{S_\tau \cap \Omega^-})} \ .\end{aligned}$$

Then Lemma 2.37 and Assumption 2.25 give

$$\begin{aligned}|(u - J_H u)^\wedge|_{H^1(\widehat{\gamma})} &\lesssim (1 + \eta_H) \Big(H_{\widehat{\tau}}^{-1} \|\widehat{u} - \widehat{q}\|_{L_2(\widehat{S_\tau \cap \Omega^-})} + |\widehat{u} - \widehat{q}|_{H^1(\widehat{S_\tau \cap \Omega^-})} \\ &\qquad + H_{\widehat{\tau}}^{\epsilon} |\widehat{u} - \widehat{q}|_{H^{1+\epsilon}(\widehat{S_\tau \cap \Omega^-})} \Big) \\ &\lesssim (1 + \eta_H) \|\widehat{u} - \widehat{q}\|_{H^2(\widehat{S_\tau \cap \Omega^-})} \ ,\end{aligned}$$

since $\operatorname{diam}((S_\tau \cap \Omega^-)^\wedge) \sim H_{\widehat{\tau}} \sim 1$. Then by Lemma 2.36 we obtain

$$|(u - J_H u)^\wedge|_{H^1(\widehat{\gamma})} \lesssim (1 + \eta_H) |\widehat{u}|_{H^2(\widehat{S_\tau \cap \Omega^-})} \ .$$

Combining this with (2.38) and (2.39) we obtain

$$\begin{aligned}|u - J_H u|_{H^1(\gamma)} &\lesssim \rho_\tau^{-1} |\det A_\tau|^{\frac{1}{2}} |(u - J_H u)^\wedge|_{H^1(\widehat{\gamma})} \\ &\lesssim \rho_\tau^{-1} |\det A_\tau|^{\frac{1}{2}} (1 + \eta_H) |\widehat{u}|_{H^2(\widehat{S_\tau \cap \Omega^-})} \\ &\lesssim \rho_\tau^{-1} |\det A_\tau|^{\frac{1}{2}} H_\tau^2 |\det A_\tau|^{-\frac{1}{2}} (1 + \eta_H) |u|_{H^2(S_\tau \cap \Omega^-)} \\ &\lesssim H_\tau (1 + \eta_H) |u|_{H^2(S_\tau \cap \Omega^-)}\end{aligned}$$

using the shape regularity of τ. □

We now modify the argument in Theorem 2.42 to bound the error in the low coefficient component of (2.55). As in the argument for Theorem 2.42, $u \notin H^2$ on all of τ and $J_H u$ samples u only in a high coefficient region around τ, therefore we cannot use the pullback to obtain a similar error estimate on τ^+.

Theorem 2.51. *Suppose $\tau \in \mathcal{T}_H^C(\Omega)$, then define $\gamma := \tau^+$ (see (2.25)). Alternatively, if $\tau \in \mathcal{T}_H^B(\Omega^+)$, then define $\gamma := \tau$. Then*

$$|u - J_H u|_{H^1(\gamma)} \lesssim H_\tau^{\frac{1}{2} - \varepsilon} (1 + \eta_H) |u|_{H^{\frac{3}{2} - \varepsilon}(S_\tau^*)} \tag{2.57}$$

where $S_\tau^* = \bigcup_{\tau' \in S_\tau} S_{\tau'}$ is the set of neighbours to S_τ.

Proof. Since $J_H q = q$ for $q \in V_H^{\mathbb{P}_1}$ we have

$$\begin{aligned}|u - J_H u|_{H^1(\gamma)} &= |(u-q) - J_H(u-q)|_{H^1(\gamma)} \\ &\leq |(u-q)|_{H^1(\gamma)} + |J_H(u-q)|_{H^1(\gamma)} \\ &\lesssim |(u-q)|_{H^1(\gamma)} + \eta_H \|u-q\|_{L_\infty(S_\tau)} \ ,\end{aligned}$$

where the last step uses Lemma 2.48 and a decomposition of the form (2.54) without the pullback, noting that $H_\gamma/\rho_\tau \lesssim 1$. Noting that S_τ is the set of neighbours to τ thus the shape regularity of $\mathcal{T}_H(\Omega)$ ensures S_τ is also shape regular ($\operatorname{diam}(S_\tau) \sim H_\tau \sim \rho_\tau$), using Lemma 2.37 we have,

$$|u - J_H u|_{H^1(\gamma)} \lesssim (1+\eta_H)\left(H_\tau^{-1}\|u-q\|_{L_2(S_\tau)} + |u-q|_{H^1(S_\tau)} + H_\tau^{\epsilon}|u-q|_{H^{1+\epsilon}(S_\tau)}\right)$$

as $\gamma \subset S_\tau$. Finally by Theorem 2.41 there exists a $q \in V_H^{\mathbb{P}_1}$ such that

$$\begin{aligned}|u - J_H u|_{H^1(\gamma)} &\lesssim (1+\eta_H)\left(H_\tau^{-1+\frac{3}{2}-\epsilon} + H_\tau^{\frac{1}{2}-\epsilon} + H_\tau^{\epsilon+\frac{3}{2}-2\epsilon}\right)|u|_{H^{\frac{3}{2}-\epsilon}(S_\tau^*)} \\ &\lesssim H_\tau^{\frac{1}{2}-\epsilon}(1+\eta_H)\,|u|_{H^{\frac{3}{2}-\epsilon}(S_\tau^*)}\end{aligned}$$

as required. □

Now, combining Theorems 2.50 and 2.51 along with (2.55) we get the following theorem.

Theorem 2.52. *Suppose $\tau \in \mathcal{T}_H^C(\Omega)$, then*

$$\begin{aligned}|u - J_H u|^2_{H^1(\tau),\alpha} \lesssim\ & \alpha^-\left(H_\tau\,(1+\eta_H)\,|u|_{H^2(S_\tau\cap\Omega^-)}\right)^2 \\ &+ \alpha^+\left(H_\tau^{\frac{1}{2}-\varepsilon}\,(1+\eta_H)\,|u|_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)}\right)^2 \ ,\end{aligned} \tag{2.58}$$

where $S_\tau^* = \bigcup_{\tau' \in S_\tau} S_{\tau'}$.

Proof. The proof follows directly from (2.55) and Theorems 2.50 and 2.51. □

2.4 A priori error bound on the whole domain

In Theorems 2.50 and 2.51 we have given error estimates on τ when τ is either a cut or border element, but we still have to give an error estimate for the remaining elements that are away from the interface. By construction of the function $J_H u$ in Definition

2.46 we have that $J_H u(x_i) = u(x_i)$ for any $x_i \in \tau$, $\tau \notin \mathcal{T}_H^C(\Omega) \cup \mathcal{T}_H^B(\Omega)$. Thus as $J_H u$ is linear on each element then

$$J_H u|_\tau \equiv I_\tau u|_\tau \tag{2.59}$$

for any element $\tau \in \mathcal{T}_H(\Omega) \backslash \left(\mathcal{T}_H^C(\Omega) \bigcup \mathcal{T}_H^B(\Omega)\right)$ where $I_\tau u$ is the conventional linear nodal interpolant. Therefore we can use a standard a priori error estimate for these elements that are away from the interface.

Proposition 2.53. *For any $\tau \in \mathcal{T}_H(\Omega) \backslash \left(\mathcal{T}_H^C(\Omega) \bigcup \mathcal{T}_H^B(\Omega)\right)$ and supposing that $\tau \subset \Omega_i$,*

$$|u - J_H u|_{H^1(\tau),\alpha} \lesssim \sqrt{\alpha_i} H_\tau |u|_{H^2(\tau)} . \tag{2.60}$$

Proof. By (2.59),

$$|u - J_H u|^2_{H^1(\tau),\alpha} = \alpha_i |u - J_H u|^2_{H^1(\tau)} = \alpha_i |u - I_\tau u|^2_{H^1(\tau)} \lesssim \alpha_i \left(H_\tau |u|_{H^2(\tau)}\right)^2$$

where the last line is a result of applying Theorem 4.4.4 in [20]. We then take the square root of both sides to get the final result. □

Now we bring all of these components together to create a robust error estimate on the whole domain. To do this we decompose the error on the whole domain into the error on the cut elements, border elements and all remaining elements. Thus for the mesh $\mathcal{T}_H(\Omega)$ we have the decomposition

$$|u - J_H u|^2_{H^1(\Omega),\alpha} = \left(\sum_{\tau \in \mathcal{T}_H^C(\Omega)} + \sum_{\tau \in \mathcal{T}_H^B(\Omega)} + \sum_{\tau \in \mathcal{T}_H^R(\Omega)} \right) |u - J_H u|^2_{H^1(\tau),\alpha} \ , \tag{2.61}$$

where

$$\mathcal{T}_H^R(\Omega) := \mathcal{T}_H(\Omega) \backslash \left(\mathcal{T}_H^C(\Omega) \bigcup \mathcal{T}_H^B(\Omega)\right) \ . \tag{2.62}$$

We bring together the results of Theorems 2.50, 2.51, 2.52 and finally Proposition 2.53 to bound each of the three sums in (2.61). Crucially in this we also use the contrast explicit regularity result from Theorem 2.22. Firstly we bound the error on the elements away from an interface in $\mathcal{T}_H^R(\Omega)$.

Theorem 2.54.

$$\sum_{\tau \in \mathcal{T}_H^R(\Omega)} |u - J_H u|^2_{H^1(\tau),\alpha} \lesssim \left(H \|f\|_{L_2(\Omega)}\right)^2 \ . \tag{2.63}$$

Proof.

$$\sum_{\tau\in\mathcal{T}_H^R(\Omega)} |u-J_Hu|^2_{H^1(\tau),\alpha} \leq \sum_{i=0}^{m} \sum_{\tau\in\mathcal{T}_H^R(\Omega)\cap\Omega_i} |u-J_Hu|^2_{H^1(\tau),\alpha}$$
$$\lesssim \sum_{i=0}^{m} \sum_{\tau\in\mathcal{T}_H^R(\Omega)\cap\Omega_i} \alpha_i \left(H_\tau |u|_{H^2(\tau)}\right)^2$$

by Proposition 2.53. We can then bound each H_τ above by H as in (2.17) and combine the parts on each inclusion to get

$$\sum_{\tau\in\mathcal{T}_H^R(\Omega)} |u-J_Hu|^2_{H^1(\tau),\alpha} \lesssim H^2 \sum_{i=0}^{m} \alpha_i \left(|u|_{H^2(\Omega_i)}\right)^2$$

to which we then apply the regularity result in Theorem 2.22 giving

$$\sum_{\tau\in\mathcal{T}_H^R(\Omega)} |u-J_Hu|^2_{H^1(\tau),\alpha} \lesssim H^2 \sum_{i=0}^{m} \frac{1}{\alpha_i} \left(\|f\|_{L_2(\Omega)}\right)^2.$$

We then recall Assumption 2.6 that $\alpha_i \geq 1$. Since there are a finite number of inclusions we obtain the final result. □

Secondly we bound the summation over the border elements $\mathcal{T}_H^B(\Omega)$. This result is slightly more difficult than the previous theorem as we have to consider the two sets of border elements, the ones on the high coefficient side of the interface and the ones on the low coefficient side. To remove u from the right hand side of our error estimates we prove the following theorem about the $H^{\frac{3}{2}-\varepsilon}$ regularity of u.

Theorem 2.55. *For the solution u of Problem 2.2 we have that if $f \in H^s(\Omega)$ for any $s > 0$ then*

$$|u|_{H^{\frac{3}{2}-\varepsilon}(\widetilde{\Omega})} \lesssim \|f\|_{H^s(\Omega)} \ , \tag{2.64}$$

for all $\varepsilon \in \left(0, \frac{1}{2}\right]$ and $\widetilde{\Omega} = (\cup_{i=1}^m \Omega_i) \cup \widetilde{\Omega}_0$ (see Theorem 2.22).

Proof. First we know that since $\alpha_i \geq 1$ for any $i = 0, ..., m$ (see Assumption 2.6) then

$$|u|^2_{H^1(\Omega)} \leq |u|^2_{H^1(\Omega),\alpha} = a_\Omega(u,u) = \int_\Omega fu \leq \|f\|_{L_2(\Omega)} \|u\|_{L_2(\Omega)} \lesssim \|f\|_{L_2(\Omega)} |u|_{H^1(\Omega)} \ ,$$

where we have referred back to (2.1) and (2.3), and then using the Cauchy-Schwarz inequality and finally the Poincaré-Friedrichs inequality. Thus

$$\|u\|_{H^1(\Omega)} \lesssim \|f\|_{L_2(\Omega)} \tag{2.65}$$

From the regularity result in Theorem 2.22 we know that $u \in H^{2+s}(\widetilde{\Omega}_0)$, $u \in H^{2+s}(\Omega_i)$ for any $s \geq 0$ and $i = 1, ..., m$. Using the Slobodeckiĭ seminorm [p74 McLean [68]] we have, for $\mu \in [0, \frac{1}{2})$,

$$\begin{aligned} |u|^2_{H^{1+\mu}(\widetilde{\Omega})} &= \int_{\widetilde{\Omega}} \int_{\widetilde{\Omega}} \frac{|Du(x) - Du(y)|^2}{|x-y|^{2+2\mu}} \; dx \; dy \\ &= \sum_{i=0}^{m} \sum_{j=0}^{m} \int_{\Omega_i} \int_{\Omega_j} \frac{|Du(x) - Du(y)|^2}{|x-y|^{2+2\mu}} \; dx \; dy \end{aligned}$$

where for this proof only we have used $\Omega_0 = \widetilde{\Omega}_0$ for ease of notation. Note that when $i = j$,

$$\int_{\Omega_i} \int_{\Omega_i} \frac{|Du(x) - Du(y)|^2}{|x-y|^{2+2\mu}} \; dx \; dy = |u|^2_{H^{1+\mu}(\Omega_i)} \leq \|u\|^2_{H^2(\Omega_i)} \lesssim \left(1 + \frac{1}{\alpha_i}\right)^2 \|f\|^2_{L_2(\Omega)} \; .$$

Then when $i \neq j$,

$$\begin{aligned} &\int_{\Omega_i} \int_{\Omega_j} \frac{|Du(x) - Du(y)|^2}{|x-y|^{2+2\mu}} \; dx \; dy \\ &\qquad \leq \left(\|Du\|^2_{L_\infty(\Omega_i)} + \|Du\|^2_{L_\infty(\Omega_j)} \right) \int_{\Omega_i} \int_{\Omega_j} \frac{1}{|x-y|^{2+2\mu}} \; dx \; dy \; . \end{aligned}$$

Using the definition of the weight $w_\mu(y)$

$$w_\mu(y) := \int_{\mathbb{R}^2 \setminus \Omega_i} \frac{1}{|x-y|^{2+2\mu}} \; dx$$

from [McLean p96 [68]] we obtain,

$$\int_{\Omega_i} \int_{\Omega_j} \frac{1}{|x-y|^{2+2\mu}} \; dx \; dy \leq \int_{\Omega_i} \left[\int_{\mathbb{R}^2 \setminus \Omega_i} \frac{1}{|x-y|^{2+2\mu}} \; dx \right] \; dy = \int_{\Omega_i} w_\mu(y) \; dy \; .$$

Now $w_\mu(y) \leq C \text{dist}\,(y, \partial\Omega_i)^{-2\mu}$ [McLean p96 [68]] so

$$\int_{\Omega_i} w_\mu(y) \; dy \leq C \int_{\Omega_i} \text{dist}\,(y, \partial\Omega_i)^{-2\mu} \cdot 1 \; dy \leq C \, \|1\|_{H^s(\Omega_i)} \; ,$$

using [McLean Lemma 3.32 [68]] and thus the double integral is bounded. We then use the Sobolev embedding theorem that states H^{2+s} embeds into W^1_∞ for any $s > 0$.

Thus, assuming each inclusion is $O(1)$ in size we obtain,

$$|u|^2_{H^{\frac{3}{2}-\varepsilon}(\widetilde{\Omega})} \lesssim \sum_{i=0}^{m}\sum_{j=0}^{m} \left(\|u\|^2_{H^1(\Omega_i)} + |u|^2_{H^2(\Omega_i)} + |u|^2_{H^{2+s}(\Omega_i)} \right) + \left(\|u\|^2_{H^1(\Omega_j)} + |u|^2_{H^2(\Omega_j)} + |u|^2_{H^{2+s}(\Omega_j)} \right)$$

Then using (2.65) and Theorem 2.22 for any seminorm of order 2 or greater we have

$$|u|^2_{H^{\frac{3}{2}-\varepsilon}(\Omega)} \lesssim \sum_{i=0}^{m} \left(\|f\|^2_{L_2(\Omega)} + \frac{1}{\alpha_i^2}\|f\|^2_{L_2(\Omega)} + \frac{1}{\alpha_i^2}\|f\|^2_{H^s(\Omega)} \right) \lesssim \|f\|^2_{H^s(\Omega)}$$

as $\alpha_i \geq 1$. □

We can now use this to obtain the following error estimate for the border elements.

Theorem 2.56.

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)} |u - J_H u|^2_{H^1(\tau),\alpha} \lesssim (1+\eta_H)^2 \left(\left(H\|f\|_{L_2(\Omega)} \right)^2 + \left(H_B^{\frac{1}{2}-\varepsilon}\|f\|_{H^s(\Omega)} \right)^2 \right) \quad (2.66)$$

for any $s > 0$ where H_B is the maximum element diameter in the set of border elements given by

$$H_B := \max_{\tau\in\mathcal{T}_H^B(\Omega)} H_\tau \ . \quad (2.67)$$

Proof. We split the sum over all the border elements into a sum over all border elements that intersect each inclusion. Note that under Assumption 2.20 each interface is an intersection of Ω_0 with some Ω_i for $i = 1, ..., m$. Therefore,

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)} |u - J_H u|^2_{H^1(\tau),\alpha} \lesssim \sum_{\substack{\tau\in\mathcal{T}_H^B(\Omega)\\ \tau\subset\Omega_0}} |u - J_H u|^2_{H^1(\tau),\alpha} + \sum_{i=1}^{m} \sum_{\substack{\tau\in\mathcal{T}_H^B(\Omega)\\ \tau\subset\Omega_i}} |u - J_H u|^2_{H^1(\tau),\alpha} \ . \quad (2.68)$$

We will consider Case I ((2.8): $\alpha_0 = 1$ on Ω_0 and $\alpha_i \geq \hat{\alpha}$ on Ω_i for $i = 1, ..., m$) and Case II ((2.9): $\alpha_0 = \hat{\alpha}$ on Ω_0 and $\alpha_i \leq K$ on Ω_i for $i = 1, ..., m$) separately. Firstly for Case I Ω_0 is the low coefficient side of each interface. Thus applying Theorem 2.51 we get

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}|u-J_Hu|^2_{H^1(\tau),\alpha}\lesssim\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}\left(H_\tau^{\frac{1}{2}-\varepsilon}\left(1+\eta_H\right)|u|_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)}\right)^2$$
$$\lesssim H_B^{1-2\varepsilon}\left(1+\eta_H\right)^2\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}|u|^2_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)}\ .$$

Recall that $S_\tau^*=\bigcup_{\tau'\in S_\tau}S_{\tau'}$ is approximately all the elements within a two element ball around τ. Next, as each S_τ^* contains a finite number of elements independent of H and by Assumption 2.20 they are sufficiently far from the boundary to be contained within $\widetilde{\Omega}_0$ then we get

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}|u|^2_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)}\lesssim|u|^2_{H^{\frac{3}{2}-\varepsilon}(\widetilde{\Omega})}\lesssim\|f\|^2_{H^s(\Omega)}\ ,$$

by Theorem 2.55 for any $s>0$, and so

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}|u-J_Hu|^2_{H^1(\tau),\alpha}\lesssim\left(H_B^{\frac{1}{2}-\varepsilon}\left(1+\eta_H\right)\|f\|_{H^s(\Omega)}\right)^2\ .$$

This bounds the first term on the right hand side of (2.68). For the second term, noting that in Case I Ω_i plays the role of Ω^-, we can use Theorem 2.50 to get

$$\sum_{i=1}^m\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}|u-J_Hu|^2_{H^1(\tau),\alpha}\lesssim\sum_{i=1}^m\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}\alpha_i\left(H_\tau\left(1+\eta_H\right)|u|_{H^2(S_\tau\cap\Omega_i)}\right)^2$$
$$\lesssim\left(H\left(1+\eta_H\right)\right)^2\sum_{i=1}^m\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}\alpha_i\left(|u|_{H^2(S_\tau\cap\Omega_i)}\right)^2\ .$$

Next, as each S_τ contains a finite number of elements bounded independent of H then we get

$$\sum_{i=1}^m\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}\alpha_i\left(|u|_{H^2(S_\tau\cap\Omega_i)}\right)^2\lesssim\sum_{i=1}^m\alpha_i|u|^2_{H^2(\Omega_i)}\lesssim\sum_{i=1}^m\frac{1}{\alpha_i}\|f\|^2_{L_2(\Omega)}$$

by Theorem 2.22. Thus as there are a finite number of inclusions we get

$$\sum_{i=1}^m\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}|u-J_Hu|^2_{H^1(\tau),\alpha}\lesssim\left(H_B\left(1+\eta_H\right)\|f\|_{L_2(\Omega)}\right)^2\ .$$

Now we consider Case II (2.9) where α_0 tends to infinity and $\alpha_i\leq K$ for $i=1,...,m$. Then Ω_0 is the "high coefficient side" and each inclusion Ω_i $(i=1,...,m)$ is the "low

coefficient side" of the interface. We apply Theorem 2.51 to get

$$\sum_{i=1}^{m}\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}|u-J_Hu|^2_{H^1(\tau),\alpha}\lesssim\sum_{i=1}^{m}\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}\alpha_i\left(H_\tau^{\frac{1}{2}-\varepsilon}\left(1+\eta_H\right)|u|_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)}\right)^2$$
$$\lesssim\left(H_B\left(1+\eta_H\right)\right)^2\sum_{i=1}^{m}\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}\alpha_i\left(|u|_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)}\right)^2$$

where H_B is as in (2.67). Next, as each S_τ^* contains a finite number of elements independent of H then we get

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}|u|^2_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)}\lesssim|u|^2_{H^{\frac{3}{2}-\varepsilon}(\widetilde{\Omega})}\lesssim\|f\|^2_{H^s(\Omega)}$$

for any $s>0$, and since we have a finite number of inclusions we get

$$\sum_{i=1}^{m}\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_i}|u-J_Hu|^2_{H^1(\tau),\alpha}\lesssim\left(H_B^{\frac{1}{2}-\varepsilon}\left(1+\eta_H\right)\|f\|^2_{H^s(\Omega)}\right)^2.$$

Now on the high coefficient side of the interfaces for Case II we can use Theorem 2.50 to get

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}|u-J_Hu|^2_{H^1(\tau),\alpha}\lesssim\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}\alpha_0\left(H_\tau\left(1+\eta_H\right)|u|_{H^2(S_\tau\cap\Omega_0)}\right)^2$$
$$\lesssim\left(H_B\left(1+\eta_H\right)\right)^2\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}\alpha_0\left(|u|_{H^2(S_\tau\cap\Omega_0)}\right)^2.$$

Next, as each S_τ contains a finite number of elements independent of H then we get

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}\alpha_0\left(|u|_{H^2(S_\tau\cap\Omega_0)}\right)^2\lesssim\alpha_0\,|u|^2_{H^2(\Omega_0)}\lesssim\frac{1}{\alpha_0}\|f\|^2_{L_2(\Omega)}$$

by Theorem 2.22. Thus as $\alpha_0\geq1$ we get

$$\sum_{\tau\in\mathcal{T}_H^B(\Omega)\cap\Omega_0}|u-J_Hu|^2_{H^1(\tau),\alpha}\lesssim\left(H_B\left(1+\eta_H\right)\|f\|_{L_2(\Omega)}\right)^2.$$

□

Lastly we bound the sum over the cut elements in (2.61). To do this we apply Theorem 2.52 and obtain the following theorem.

Theorem 2.57.

$$\sum_{\tau\in\mathcal{T}_H^C(\Omega)} |u - J_H u|^2_{H^1(\tau),\alpha} \lesssim (1+\eta_H)^2 \left(H_C^2 \|f\|^2_{L_2(\Omega)} + H_C^{1-2\varepsilon} \|f\|^2_{H^s(\Omega)} \right) \tag{2.69}$$

for any $s > 0$, where H_C is the maximum element diameter in the set of cut elements given by

$$H_C := \max_{\mathcal{T}_H^C(\Omega)} H_\tau \ . \tag{2.70}$$

Proof. Using Theorem 2.52 we have

$$\sum_{\tau\in\mathcal{T}_H^C(\Omega)} |u - J_H u|^2_{H^1(\tau),\alpha} \lesssim (1+\eta_H)^2 \sum_{\tau\in\mathcal{T}_H^C(\Omega)} \alpha^- \left(H_\tau |u|_{H^2(S_\tau\cap\Omega^-)} \right)^2 + \alpha^+ \left(H_\tau^{\frac{1}{2}-\varepsilon} |u|_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)} \right)^2 ,$$

where α^- is the coefficient on the "high coefficient side" τ^- and α^+ is the value on the "low coefficient side" (see Definition 2.23).

Consider first the high coefficient side. Then as S_τ has a finite number of neighbours independent of H we know that

$$\begin{aligned}
\sum_{\tau\in\mathcal{T}_H^C(\Omega)} \alpha^- \left(H_\tau |u|_{H^2(S_\tau\cap\Omega^-)} \right)^2 &\lesssim H_C^2 \sum_{\tau\in\mathcal{T}_H^C(\Omega)} \alpha^- |u|^2_{H^2(S_\tau\cap\Omega^-)} \\
&\lesssim H_C^2 \sum_{i=0}^{m} \alpha_i |u|^2_{H^2(\Omega_i)} \\
&\lesssim H_C^2 \sum_{i=0}^{m} \frac{1}{\alpha_i} \|f\|^2_{L_2(\Omega)} \\
&\lesssim H_C^2 \|f\|^2_{L_2(\Omega)}
\end{aligned}$$

using the regularity result in Theorem 2.22 and that $\alpha_i \geq 1$ for any $i = 0, ..., m$. It applies in both Case I (2.8) with Ω^- as each inclusion Ω_i for $i = 1, ..., m$ and $\Omega^+ = \Omega_0$, and also Case II (2.9) with $\Omega^- = \Omega_0$ and Ω^+ as each inclusion Ω_i for $i = 1, ..., m$.

On the low coefficient side Ω^+ we get

$$\sum_{\tau\in\mathcal{T}_H^C(\Omega)} \alpha^+ \left(H_\tau^{\frac{1}{2}-\varepsilon} |u|_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)} \right)^2 \lesssim H_C^{1-2\varepsilon} \sum_{\tau\in\mathcal{T}_H^C(\Omega)} \alpha^+ |u|^2_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)}$$

and since there are a finite number of elements in S_τ^* independent of H we have

$$\sum_{\tau\in\mathcal{T}_H^C(\Omega)} \alpha^+ \left(H_\tau^{\frac{1}{2}-\varepsilon} |u|_{H^{\frac{3}{2}-\varepsilon}(S_\tau^*)} \right)^2 \lesssim H_C^{1-2\varepsilon} \max\{1,K\} |u|^2_{H^{\frac{3}{2}-\varepsilon}(\widetilde{\Omega})} \lesssim H_C^{1-2\varepsilon} \|f\|^2_{H^s(\Omega)}$$

because for either Case I or II α^+ is bounded above by $\max\{1,K\}$ (see (2.8) and (2.9)). □

The results of Theorems 2.54, 2.56 and 2.57 can then be substituted into (2.61) to give the final robust error bound on the whole domain.

Theorem 2.58. *The finite element error $u-u_H$ in the energy norm is bounded by*

$$|u-u_H|_{H^1(\Omega),\alpha} \lesssim (1+\eta_H) H^{\frac{1}{2}-\epsilon} \left(H^{1+2\epsilon} \|f\|^2_{L_2(\Omega)} + \delta_H^{1-2\epsilon} \|f\|^2_{H^s(\Omega)} \right)^{\frac{1}{2}} \tag{2.71}$$

for any $s>0$, where

$$\delta_H = \frac{\max\{H_B,H_C\}}{H} \ . \tag{2.72}$$

Proof. The proof follows from substituting (2.69), (2.66) and (2.63) into (2.61) and using the bound $\max\{H_B,H_C\} \le \delta_H H$. □

This leads to the following corollary that shows robust $O(H)$ convergence in the energy norm can be restored given sufficient refinement of the mesh around the interfaces.

Corollary 2.59. *Suppose $\mathcal{T}_H(\Omega)$ is a quasi-uniform mesh with sufficient refinement of elements around the interfaces such that*

$$\max\{H_B,H_C\} \le H^{\frac{2}{1-2\epsilon}}$$

then

$$|u-u_H|_{H^1(\Omega),\alpha} \lesssim H\,(1+\eta_H) \left(\|f\|^2_{L_2(\Omega)} + \|f\|^2_{H^s(\Omega)} \right)^{\frac{1}{2}} \ . \tag{2.73}$$

Proof. The proof follows from Theorem 2.58 and that $\max\{H_B,H_C\} \lesssim H^{\frac{2}{1-2\epsilon}}$ implies

$$\delta_H^{\frac{1}{2}-\epsilon} \lesssim \left(H^{\frac{2}{1-2\epsilon}-1} \right)^{\frac{1}{2}-\epsilon} = \left(H^{\frac{1+2\epsilon}{1-2\epsilon}} \right)^{\frac{1}{2}-\epsilon} = H^{\frac{1}{2}+\epsilon} \ .$$

□

We then relate these finite element error estimates in the energy norm back to the L_2 norm using a non-standard variant of the usual duality argument.

Theorem 2.60. *The finite element error $u - u_H$ in the L_2 norm is bounded by*

$$\|u - u_H\|_{L_2(\Omega)} \lesssim (1 + \eta_H)^2 H^{1-2\epsilon} \|f\|_{H^s(\Omega)} \ . \tag{2.74}$$

Proof. Consider the dual problem where $w \in H_0^1(\Omega)$ and $w_H \in V_H^{\mathbb{P}_1}$ solve

$$\begin{aligned} a_\Omega(w, v) &= (u - u_H, v) \quad \text{for all } v \in H_0^1(\Omega) \\ a_\Omega(w_H, v_H) &= (u - u_H, v_H) \quad \text{for all } v_H \in V_H^{\mathbb{P}_1} \end{aligned}$$

respectively. Then by Theorem 2.58 where $\epsilon > 0$ we have

$$|w - w_H|_{H^1(\Omega),\alpha} \lesssim \left((1 + \eta_H) H^{\frac{1}{2}-\epsilon}\right) \left(H^{1+2\epsilon}(1 + \delta_H)^2 + \delta_H^{1-2\epsilon}\right)^{\frac{1}{2}} \|u - u_H\|_{H^s(\Omega)}$$

for any $s > 0$. From (2.72) we know that $\delta_H \leq 1$ and define $\rho(H) := (1 + \eta_H) H^{\frac{1}{2}-\epsilon}$, then for $s = \frac{1}{2}$

$$\begin{aligned} |w - w_H|_{H^1(\Omega),\alpha} &\lesssim \rho(H) \|u - u_H\|_{H^{\frac{1}{2}}(\Omega)} \\ &\lesssim \rho(H) \left(|u - u_H|_{H^1(\Omega)} \|u - u_H\|_{L_2(\Omega)} + \|u - u_H\|_{L_2(\Omega)}^2\right)^{\frac{1}{2}} \end{aligned}$$

by the interpolation theorem for $H^{\frac{1}{2}}(\Omega)$. By the arithmetic-geometric mean inequality ($ab \leq a^2/2 + b^2/2$) we get

$$\begin{aligned} |w - w_H|_{H^1(\Omega),\alpha} &\lesssim \rho(H) \left(\rho(H)^2 |u - u_H|_{H^1(\Omega)}^2 + \|u - u_H\|_{L_2(\Omega)}^2\right)^{\frac{1}{2}} \\ &\lesssim \rho(H)^2 |u - u_H|_{H^1(\Omega)} + \rho(H) \|u - u_H\|_{L_2(\Omega)} \ . \end{aligned} \tag{2.75}$$

However we also know that

$$\begin{aligned} \|u - u_H\|_{L_2(\Omega)}^2 = a_\Omega(w, u - u_H) &= a_\Omega(w - w_H, u - u_H) \\ &\lesssim |w - w_H|_{H^1(\Omega),\alpha} |u - u_H|_{H^1(\Omega),\alpha} \ . \end{aligned}$$

Combining this with (2.75) we get

$$\|u - u_H\|_{L_2(\Omega)}^2 \lesssim \rho(H)^2 |u - u_H|_{H^1(\Omega)}^2 + \rho(H) \|u - u_H\|_{L_2(\Omega)} |u - u_H|_{H^1(\Omega),\alpha}$$

which by the arithmetic-geometric mean inequality again gives

$$\rho(H)\left\|u-u_H\right\|_{L_2(\Omega)}\left|u-u_H\right|_{H^1(\Omega),\alpha}\lesssim\frac{1}{2}\rho(H)^2\left|u-u_H\right|^2_{H^1(\Omega),\alpha}+\frac{1}{2}\left\|u-u_H\right\|^2_{L_2(\Omega)}\ .$$

Thus

$$\frac{1}{2}\left\|u-u_H\right\|^2_{L_2(\Omega)}\lesssim\rho(H)^2\left|u-u_H\right|^2_{H^1(\Omega)}\ .$$

Finally using Theorem 2.58 again we get

$$\left\|u-u_H\right\|_{L_2(\Omega)}\lesssim\rho(H)^2\left\|f\right\|_{H^s(\Omega)}=(1+\eta_H)^2H^{1-2\epsilon}\left\|f\right\|_{H^s(\Omega)}\ .$$

□

We also obtain a similar bound for a quasi-uniform mesh with sufficient refinement around the interfaces by using Corollary 2.59 instead of Theorem 2.58 in the proof of the L_2 error estimate. For a mesh with sufficient refinement around the interfaces we obtain

$$\left\|u-u_H\right\|_{L_2(\Omega)}\lesssim H^2\left(1+\eta_H\right)^2\left\|f\right\|_{H^s(\Omega)}\ . \tag{2.76}$$

We note that similar results in terms of powers of H are obtained by Li, Melenk, Wohlmuth and Zou in [59] but it has no estimates that are explicit in the contrast.

Remark 2.61. *We remark that it may be possible to extend these results to $\Omega\subset\mathbb{R}^3$ but many of the proofs will require a new method. Many of the results in this chapter rely on the embedding of $H^{1+\varepsilon}(\Omega)$ into $L_\infty(\Omega)$ for $\varepsilon>0$. In 3D $H^{\frac{3}{2}+\varepsilon}(\Omega)$ embeds into $L_\infty(\Omega)$ but the solution u is only in $H^{\frac{3}{2}-\varepsilon}(\Omega)$ and therefore the same techniques cannot be used. Further study, both numerical and analytical, are required for a 3D result.*

2.5 Summary

In this chapter we have given a detailed description of the high contrast elliptic interface problem and shown how to formulate the finite element approximation of the solution using the standard finite element method.

We have shown that, under certain assumptions, we obtain a new finite element error estimate in the energy norm and L_2 norm that is independent of the contrast in the coefficient $\mathcal{A}(x)$. This was done using a simplified proof for a single element that was

cut by the interface and then a technical argument to adapt the argument to produce a conforming approximation across the domain Ω.

While this chapter indeed shows contrast independence in the finite element error it also confirms the lower rate of convergence that is observed (see Section 4.6), being $O(H^{\frac{1}{2}-\epsilon})$ in the energy norm where $O(H)$ is expected if the coefficient $\mathcal{A}$ were smooth. In the following chapters we explore methods to restore this improved rate whilst retaining contrast independent finite element errors.

3

Extensions to the Multiscale Finite Element Method

In Chapter 2 we presented an a priori error estimate for the standard FEM applied to the scaled interface problem (Problem 2.2). This is robust with respect to the contrast parameter $\hat{\alpha}$ but is only $O(H^{\frac{1}{2}-\epsilon})$ in the energy norm. The standard finite element method uses the space of continuous piecewise linear functions $V_H^{\mathbb{P}_1}$ but we demonstrated how, with sufficient refinement of the mesh (effectively resolving the interfaces), we can restore $O(H)$ convergence in the energy norm. The problem with having to resolve the interfaces with the mesh is that it increases the amount of computational work to be done. In fact it may not be easy to resolve the interfaces especially if they are very complicated. For example in Chapter 5 we will introduce the shape optimisation problem where inclusions of all shapes and sizes are introduced into a solid material.

Now we present a different approach where, instead of approximating in $V_H^{\mathbb{P}_1}$, we use a better space of multiscale functions $V_H^{\text{MS}} \subset H_0^1(\Omega)$. The multiscale finite element solution u_H^{MS} found from solving the finite element problem using V_H^{MS} (Problem 2.11) should produce an error estimate that is $O(H)$ in the energy norm independent of the contrast parameter $\hat{\alpha}$ and without extra refinement of elements near the interfaces provided V_H^{MS} is properly chosen.

The space V_H^{MS} is the span of a set of multiscale basis functions $\{\Phi_i^{\text{MS}}\}$ which are defined for each node $n_i \in \mathcal{N}(\mathcal{T}_H(\Omega))$ of a coarse mesh $\mathcal{T}_H(\Omega)$. The basis functions Φ_i^{MS} attempt to incorporate the fine scale features of the permeability field $\mathcal{A}(x)$ in (1.1), or more specifically, in the case of interface problems like Problem 2.2, they incorporate the discontinuous gradient of the solution u that results from a discontinuous permeability field $\alpha(x)$ running through the interior of an element.

In this chapter we will review the work of Chu, Graham and Hou in [27] that presents

multiscale finite element error estimates that, under certain conditions, are independent of the contrast parameter $\hat{\alpha}$ and that allows the interface to pass through the interior of an element. They achieve this by solving local homogeneous problems on each cut element to obtain the multiscale basis functions Φ_i^{MS} subject to suitable local artificial boundary conditions. Crucially their proofs make no appeal to homogenisation theory, unlike a lot of previous work (as discussed in Section 1.2.2). We will discuss their method further in Section 3.1 with the aim of providing clearer insight into the analysis. We will show how their construction of local boundary conditions is simple and how implementation is easy, hence ideal for use in a practical multiscale finite element code. It is the analysis behind their robust finite element error estimate that makes [27] difficult. We will try to distill the key elements of the proof and refer to [27] for the technical detail.

In Section 3.1.1 we start by restating a key idea for proving a robust a priori finite element error for multiscale problems. This idea was introduced in [27] for high contrast elliptic interface problems but we interpret it here in the more general case of high contrast multiscale elliptic problems. Our interpretation also does not require the nodal interpolant but instead applies to a general $v_H \in V_H^{\text{MS}}$. We also give a new insight into [27] in Section 3.1.1 by first describing the steps of their proof and leaving the technical detail for later sections. This new insight seeks to emphasise the method and how easy it is to implement. In Sections 3.1.3 and 3.1.4 we walk through the analysis of [27] in more detail to provide an easier understanding of the MsFEM analysis but also in Section 3.1.5 we provide a new generalisation of the interior error result [Lemma 3.15 [27]] that does not rely on using the nodal interpolant.

In both Chapter 2 and the present chapter we consider the scaled interface problem, Problem 2.2, where the permeability field $\alpha(x) \geq 1$ and the contrast parameter $\hat{\alpha}$ may be unboundedly large. However we commented in Remark 2.8 on how Problem 2.2 can be related to the case when the coefficient $\mathcal{A}(x)$ tends to zero and by proving a lower bound on the H^2 seminorm of u we can obtain a relative error estimate. We will explore this in Section 3.2 by first introducing a new generalisation of the proof of the regularity result from [27] (restated in this thesis as Theorem 2.22) to multiple inclusions instead of just a single inclusion. We will then utilise this proof to show a lower bound on the H^2 seminorm of u and then finally prove a relative error estimate. The relative error estimate is new.

3.1 The Multiscale Finite Element Method of Graham, Chu and Hou

The version of the Multiscale Finite Element Method (MsFEM) introduced in [27] provides a new algorithm and error analysis for the high-contrast elliptic interface problem that we introduced in Chapter 2: Find $u \in H_0^1(\Omega)$ such that

$$a_\Omega(u,v) := \int_\Omega \nabla u \cdot \alpha \nabla v \ \ dx = \int_\Omega fv \ \ dx \qquad \text{for any } v \in H_0^1(\Omega) \tag{3.1}$$

(see Problem 2.1). The analysis of [27] was restricted to the situtation where $\alpha(x) \geq 1$ and is constant on a finite number of inclusions within the domain Ω (see Definition 2.4 and Assumption 2.6) and focussed on the high contrast cases where $\hat{\alpha}$ is large (Case I (2.8) and Case II (2.9)).

The method in [27] involves creating nodal multiscale basis functions on a (coarse) quasiuniform triangular mesh $\mathcal{T}_H(\Omega)$. The basis functions coincide with the linear hat functions used to span $V_H^{\mathbb{P}_1}$ on elements where α is constant. Otherwise they are precomputed by solving a local homogeneous version of (3.1) subject to artificial boundary conditions. The resulting basis functions can then be used to define a multiscale finite element solution u_H^{MS} by the Galerkin method. The choice of boundary condition is key to proving a robust finite element error estimate. In [27], under certain conditions, error estimates of the form

$$|u - u_H^{\text{MS}}|_{H^1(\Omega),\alpha} \lesssim H \left[H \, |f|^2_{H^{\frac{1}{2}}(\Omega)} + \|f\|^2_{L_2(\Omega)} \right]^{\frac{1}{2}} \tag{3.2}$$

are proved, where the hidden constant is again independent of H and the contrast parameter $\hat{\alpha}$ (see Notation 2.14). This is a big improvement when compared to our result in Theorem 2.58 for the standard finite element method which also showed independence from $\hat{\alpha}$ but was only $O(H^{\frac{1}{2}-\epsilon})$ when the mesh does not resolve the interface. In [27] a non-standard duality argument is also used to show

$$\|u - u_H^{\text{MS}}\|_{L_2(\Omega)} \lesssim H^2 \left[H \, |f|^2_{H^{\frac{1}{2}}(\Omega)} + \|f\|^2_{L_2(\Omega)} \right]^{\frac{1}{2}} . \tag{3.3}$$

The techniques from this duality argument also led to the L_2 finite element error estimate in Thereom 2.60 for the standard finite element method. The disadvantage of MsFEM compared to the standard FEM is the need for the solution of local subgrid problems on elements that have the interface running through their interior and a slightly worse dependence on f.

3.1.1 A key idea behind the multiscale finite element method

A key idea in MsFEM is to solve a local subgrid problem to obtain multiscale basis functions that form a better approximation space. For this we require a local version of $a_\Omega(\cdot,\cdot)$. For any measurable $D \subset \Omega$ we write

$$a_D(v,w) = \int_D \nabla v \cdot \alpha \nabla w \;\; dx \; . \tag{3.4}$$

Then for any triangular element $\tau \in \mathcal{T}_H(\Omega)$ with the three nodes $x_i^\tau, x_j^\tau, x_k^\tau \in \mathcal{N}(\mathcal{T}_H(\Omega))$ we shall construct nodal basis functions Φ_p^{MS} $(p = i,j,k)$ whose restriction $\Phi_{p,\tau}^{\mathrm{MS}}$ to τ must solve the following subgrid problem.

Problem 3.1 (Subgrid Problem)**.** *Find $\Phi_{p,\tau}^{MS} \in H^1(\tau)$ such that*

$$a_\tau(\Phi_{p,\tau}^{MS}, v) = 0 \qquad \textit{for all } v \in H_0^1(\tau) \; , \tag{3.5}$$

subject to a suitable boundary condition

$$\Phi_{p,\tau}^{MS} = \phi_{p,\tau} \qquad \textit{on } \partial\tau \; , \tag{3.6}$$

where $\phi_{p,\tau} \in C(\partial\tau)$, $\phi_{p,\tau}(x_q^\tau) = \delta_{pq}$ for $p,q \in \{i,j,k\}$ and $\sum_{p\in\{i,j,k\}} \phi_{p,\tau} = 1$ on $\partial\tau$.

In general the boundary conditions $\phi_{p,\tau}$ have to be prescribed and the local subgrid problem solved. However, in the case when α is constant on an element and linear boundary conditions $\phi_{p,\tau}$ are used, then $\Phi_{p,\tau}^{\mathrm{MS}}$ is just the usual linear hat function restricted to τ. The local boundary conditions will be constructed so that they are continuous across element edges and so the space

$$V_H^{\mathrm{MS}} := \operatorname{span}\left\{ \Phi_p^{\mathrm{MS}} \mid x_p \in \mathcal{N}(\mathcal{T}_H(\Omega)) \right\} \subseteq H_0^1(\Omega) \; . \tag{3.7}$$

This means that the MsFEM is a conforming method. Using $V_H = V_H^{\mathrm{MS}}$ in the finite element problem (Problem 2.11) gives a multiscale finite element approximation u_H^{MS} which satisfies

$$a_\Omega(u_H^{\mathrm{MS}}, v_H^{\mathrm{MS}}) = \int_\Omega \nabla u_H^{\mathrm{MS}} \cdot \alpha \nabla v_H^{\mathrm{MS}} \;\; dx = \int_\Omega f v_H^{\mathrm{MS}} \;\; dx = L_\Omega(v_H^{\mathrm{MS}}) \quad \text{for all } v_H^{\mathrm{MS}} \in V_H^{\mathrm{MS}} \; . \tag{3.8}$$

The finite element error can be trivially bounded using Galerkin orthogonality (Lemma 2.21):

$$|u - u_H^{\mathrm{MS}}|_{H^1(\Omega),\alpha} \leq |u - v_H^{\mathrm{MS}}|_{H^1(\Omega),\alpha} \tag{3.9}$$

for any $v_H^{\mathrm{MS}} \in V_H^{\mathrm{MS}}$. To bound the right hand side of (3.9) Chu, Graham and Hou introduced the following elementary lemma.

Lemma 3.2. *Suppose D is a Lipschitz subdomain of Ω and suppose that $v \in H^1(D)$ satisfies*

$$a_D(v, w) = \int_D fw \qquad \text{for any } w \in H_0^1(D) \ .$$

Then for any $\tilde{v} \in H^1(D)$ such that the trace of $\tilde{v} - v$ vanishes on ∂D,

$$|v|_{H^1(D),\alpha} \leq |\tilde{v}|_{H^1(D),\alpha} + C\,diam(D)\,\|f\|_{L_2(D)} \tag{3.10}$$

where C is independent of v, $\tilde{v}$, the diameter of D and $\hat{\alpha}$.

Proof. Let v^* be the unique solution of the problem

$$a_D(v^*, w) = 0 \qquad \text{for any } w \in H_0^1(D) \tag{3.11}$$

such that the trace of $v^* - v$ vanishes on ∂D. Then $v - v^* \in H_0^1(D)$ and

$$a_D(v - v^*, w) = \int_D fw \ \ dx \qquad \text{for any } w \in H_0^1(D) \ .$$

Therefore

$$\begin{aligned} |v - v^*|^2_{H^1(D),\alpha} = a_D(v - v^*, v - v^*) &= \int_D f(v - v^*) \ \ dx \\ &\leq \|f\|_{L_2(D)}\,\|v - v^*\|_{L_2(D)} \\ &\leq C\mathrm{diam}(D)\,\|f\|_{L_2(D)}\,|v - v^*|_{H^1(D),\alpha} \ , \end{aligned}$$

where the last step uses the Poincaré-Friedrichs inequality and the assumption that $\alpha \geq 1$. After dividing both sides by $|v - v^*|_{H^1(D),\alpha}$ and using the inverse triangle inequality we get

$$|v|_{H^1(D),\alpha} \leq |v^*|_{H^1(D),\alpha} + C\mathrm{diam}(D)\,\|f\|_{L_2(D)} \ .$$

However (3.11) implies minimality of the energy norm of v^* so $|v^*|_{H^1(D),\alpha} \leq |\tilde{v}|_{H^1(D),\alpha}$ for all $\tilde{v}$ satisfying the same boundary conditions as v and the result follows. □

Recalling (3.5), (3.7) we note that each of the basis functions satisfies (3.5) and since any function $v_H^{\mathrm{MS}} \in V_H^{\mathrm{MS}}$ is a weighted sum of these basis functions then the local error

$E_H^{\mathrm{MS}} := u - v_H^{\mathrm{MS}}$ satisfies

$$a_\tau(E_H^{\mathrm{MS}}, w) = \int_\tau f w \ \ dx \qquad \text{for any } w \in H_0^1(\tau) \tag{3.12}$$

for any element $\tau \in \mathcal{T}_H(\Omega)$. This means we can apply Lemma 3.2 to E_H^{MS} to obtain the following theorem.

Theorem 3.3.

$$|E_H^{MS}|_{H^1(\tau),\alpha} \leq \left|\widetilde{E}_H^{MS}\right|_{H^1(\tau),\alpha} + CH_\tau \left\|f\right\|_{L_2(\tau)} \ , \tag{3.13}$$

and

$$|u - u_H^{MS}|_{H^1(\Omega),\alpha} \leq C \left[\sum_{\tau\in\mathcal{T}_H(\Omega)} \left(\left|\tilde{E}_H^{MS}\right|^2_{H^1(\tau),\alpha} + H_\tau^2 \left\|f\right\|^2_{L_2(\tau)}\right)\right]^{\frac{1}{2}} \ ,$$

where $\widetilde{E}_H^{MS}$ is any function whose trace coincides with the trace of E_H^{MS} on $\partial\tau$ and C is a generic constant independent of $\mathcal{T}_H(\Omega)$, f, u and α.

Proof. As E_H^{MS} satisfies (3.12) then Lemma 3.2 immediately gives (3.13). Next we note that

$$\begin{aligned}
|u - u_H^{\mathrm{MS}}|^2_{H^1(\Omega),\alpha} &\leq |u - v_H^{\mathrm{MS}}|^2_{H^1(\Omega),\alpha} \\
&= \sum_{\tau\in\mathcal{T}_H(\Omega)} |u - v_H^{\mathrm{MS}}|^2_{H^1(\tau),\alpha} \\
&= \sum_{\tau\in\mathcal{T}_H(\Omega)} |E_H^{\mathrm{MS}}|^2_{H^1(\tau),\alpha} \\
&\leq \sum_{\tau\in\mathcal{T}_H(\Omega)} \left(\left|\tilde{E}_H^{\mathrm{MS}}\right|_{H^1(\tau),\alpha} + CH_\tau \left\|f\right\|_{L_2(\tau)}\right)^2 \\
&\leq C \left[\sum_{\tau\in\mathcal{T}_H(\Omega)} \left(\left|\tilde{E}_H^{\mathrm{MS}}\right|^2_{H^1(\tau),\alpha} + H_\tau^2 \left\|f\right\|^2_{L_2(\tau)}\right)\right] \ ,
\end{aligned}$$

and then the result follows by taking the square root of both sides. □

This theorem shows us that if we can construct local boundary conditions so that the error E_H^{MS} on each $\partial\tau$ has an extension $\widetilde{E}_H^{\mathrm{MS}}$ into τ satisfying

$$\sum_{\tau\in\mathcal{T}_H(\Omega)} \left|\tilde{E}_H^{\mathrm{MS}}\right|^2_{H^1(\tau),\alpha} \leq CH^2 \tag{3.14}$$

where C may depend on some norm of f but not $\hat{\alpha}$ then we obtain a robust optimal error estimate. This is the fundamental idea to proving robust finite element error estimates for multiscale methods applied to high contrast heterogeneous elliptic problems of the form (1.1). Finding artificial local boundary conditions for general problems of the form (1.1) which have the required error extension is a large and difficult problem. The high contrast elliptic interface problem provides a more tractable problem and as Chu, Graham and Hou have shown in [27] it is possible, but not trivial, to define a local boundary condition such that (3.14) holds.

Remark 3.4. *We note that in [27] Chu, Graham and Hou specifically use the usual nodal interpolation operator $I_H^{MS}u$ given by*

$$I_H^{MS}u := \sum_{x_p \in \mathcal{N}(\mathcal{T}_H(\Omega))} u(x_p)\Phi_p^{MS}$$

as their choice of $v_H^{MS} \in V_H^{MS}$ in (3.9).

In [27] the authors first considered a simple application where each inclusion Ω_i is completely contained within an element τ_i. They prove that Theorem 3.3 holds by using linear boundary conditions for the subgrid problems, although, the constant C is dependent on the ratio H_{τ_i}/ϵ_i where ϵ_i is the minimum distance of Ω_i to the boundary $\partial\tau_i$. This is rather like the resonance error identified in the earlier theory of MsFEM for homogenisation problems [50] (see Seciton 1.2.2). We mentioned earlier that if the interface does not intersect an element then we also use linear boundary conditions for $\phi_{p,\tau}$ and standard finite element theory proves Theorem 3.3. We refer to [27] for the proof of Theorem 3.3 when the inclusion is inside an element and move on to consider the more practical case when the interface passes through an element.

To motivate the following sections we give an overview of the analysis in [27]. The overall idea is to show that there exists an extension $\widetilde{E}_H^{\mathrm{MS}}$ to $E_H^{\mathrm{MS}} = u - I_H^{\mathrm{MS}}u$ such that (3.14) holds. To do this the analysis in [27] uses the following ideas.

1. Under certain conditions an extension $\widetilde{E}_H^{\mathrm{MS}}$ can be constructed such that the energy norm error depends only on the error of E_H^{MS} on the boundary of τ, i.e.

$$\left|\tilde{E}_H^{\mathrm{MS}}\right|^2_{H^1(\tau),\alpha} \lesssim H_\tau^2 \left(\hat{\alpha}^2 \max_{i=1,2,3} \|D_{e_i} E_H^{\mathrm{MS}}\|^2_{L_\infty(e_i \cap \tau^-)} + \max_{i=1,2,3} \|D_{e_i} E_H^{\mathrm{MS}}\|^2_{L_\infty(e_i \cap \tau^+)}\right) \tag{3.15}$$

   where $\mathbf{e}_i$ for $i = 1, 2, 3$ are the edges of τ. We give the proof of this in Theorem 3.19, note that the proof in this thesis is a generalisation of that given in Lemma 3.15 of [27].

2. Therefore it is important to have coefficient parameter robust edge errors. With this in mind, [27] defined an algorithm (Algorithm 3.6 in this thesis) for constructing local boundary conditions $\phi_{p,\tau}$ through the solution of a small linear system. This leads to $I_H^{\mathrm{MS}} u = \sum_{p=1}^{3} u(x_p^\tau)\phi_{p,\tau}$ on $\partial\tau$.

3. They prove that these local boundary conditions $\phi_{p,\tau}$ lead to the estimate:

$$\begin{aligned}&\max_{i=1,2,3}\left\{\hat{\alpha}\left\|D_{e_i}E_H^{\mathrm{MS}}\right\|_{L_\infty(e_i\cap\tau^-)}\ ,\ \left\|D_{e_i}E_H^{\mathrm{MS}}\right\|_{L_\infty(e_i\cap\tau^+)}\right\}\\&\qquad\lesssim H_\tau^{1/2}\max_{\substack{i=1,2,3\\|k|=1}}\left[\hat{\alpha}^2\left\|D^k D_{e_i}u\right\|^2_{L_2(e_i\cap\tau^-)}+\left\|D^k D_{e_i}u\right\|^2_{L_2(e_i\cap\tau^+)}\right]^{\frac{1}{2}}\ .\end{aligned}\tag{3.16}$$

   The above bound essentially shows that the error on the edges is still of optimal order despite u having a jumping gradient and still with the dependence on $\hat{\alpha}$ explicitly stated. This is given in Theorem 3.16.

4. With the motivation of using the regularity result in Theorem 2.22, which proves the Sobolev seminorms on each inclusion are $O(\alpha_i^{-1})$, the right hand side of the previous bound is extended to the interior of the domain to give:

$$\begin{aligned}&\max_{i=1,2,3}\left\{\hat{\alpha}\left\|D_{e_i}E_H^{\mathrm{MS}}\right\|_{L_\infty(e_i\cap\tau^-)}\ ,\ \left\|D_{e_i}E_H^{\mathrm{MS}}\right\|_{L_\infty(e_i\cap\tau^+)}\right\}\\&\qquad\lesssim\left[\hat{\alpha}^2\left(|u|^2_{H^2(\tau^-)}+H_\tau|u|^2_{H^{5/2}(\tau^-)}\right)+\left(|u|^2_{H^2(\tau^+)}+H_\tau|u|^2_{H^{5/2}(\tau^+)}\right)\right]^{\frac{1}{2}}\ .\end{aligned}\tag{3.17}$$

5. Combining (3.17) with (3.15) and summing over all elements gives

$$\begin{aligned}&\sum_{\tau\in\mathcal{T}_H(\Omega)}\left|\tilde{E}_H^{\mathrm{MS}}\right|^2_{H^1(\tau),\alpha}\\&\lesssim H_\tau^2\alpha_0\left(|u|^2_{H^2(\widetilde{\Omega_0})}+H_\tau|u|^2_{H^{5/2}(\widetilde{\Omega_0})}\right)+\sum_{i=1}^{m}H_\tau^2\alpha_i\left(|u|^2_{H^2(\Omega_i)}+H_\tau|u|^2_{H^{5/2}(\Omega_i)}\right)\ .\end{aligned}\tag{3.18}$$

6. Using Theorem 2.22 this is bounded by

$$\sum_{\tau\in\mathcal{T}_H(\Omega)}\left|\tilde{E}_H^{\mathrm{MS}}\right|^2_{H^1(\tau),\alpha}\lesssim H^2\left(\|f\|^2_{L_2(\Omega)}+H|f|^2_{H^{1/2}(\Omega)}\right)\tag{3.19}$$

   and thus (3.14) holds. We sustitute this back into Theorem 3.3 to obtain a robust finite element error estimate (see Theorem 3.22).

3.1.2 An artificial local boundary condition for elements that intersect inclusions

We now present the local boundary condition $\phi_{p,\tau}$ to the subgrid problem (3.6) for elements τ that straddle the interface between inclusions as given by Chu, Graham and Hou in [27] (see Figure 3-1). Recall the notion of a cut element $\tau \in \mathcal{T}_H^C(\Omega)$ (Definition 2.18) and the corresponding regions τ^-, τ^+ with high and low coefficient respectively (see (2.25)). The method requires the following assumption.

Assumption 3.5. *Given $\tau \in \mathcal{T}_H^C(\Omega)$, we label the nodes x_1^τ, x_2^τ, x_3^τ of τ in such a way that x_3^τ is in τ^- and assume Γ intersects $\partial\tau$ at only two points $y_i = \Gamma \cap \mathbf{e}_i$ where $\mathbf{e}_i$ is the unit vector from x_3^τ in the direction $\overline{x_3^\tau x_i^\tau}$ for $i = 1, 2$. Also let β denote the angle of τ subtended at x_3^τ.*

Let r_i^- and r_i^+ denote the length of the line segments $\mathbf{e}_i \cap \tau^-$ and $\mathbf{e}_i \cap \tau^+$ respectively. Assume there exist constants $0 \leq \underline{R} \leq \overline{R} \leq 1$ and $0 < B < \pi$ such that

$$\underline{R} H_\tau \leq \min\left\{r_i^-, r_i^+\right\} \leq \max\left\{r_i^-, r_i^+\right\} \leq \overline{R} H_\tau \quad \text{for } i = 1, 2 \quad \text{and} \quad B \leq \beta \leq \pi - B.$$

For $i = 1, 2$ let $\mathbf{n}_i$ be the unit normal to Γ at the point y_i, specified to be outward from τ^-, and let $\mathbf{t}_i$ the corresponding unit tangent at y_i. Then define $\theta_i \in (-\pi/2, \pi/2)$ to be the unique angle such that

$$\mathbf{e}_i = \cos\theta_i \mathbf{n}_i + \sin\theta_i \mathbf{t}_i \ . \tag{3.20}$$

Then we also assume that neither of the edges $\mathbf{e}_i$ are tangential to Γ. So there exists a $T > 0$ such that

$$|\theta_i| \leq \pi/2 - T.$$

A typical configuration of how an element intersects the interface such that it satisfies Assumption 3.5 is shown in Figure 3-1. Note that while an element may intersect the interface in a different way, it is always possible to refine the mesh so that this assumption holds.

Before we review the analysis from [27] we first state the artificial local boundary condition. In the next section we will explain where it comes from. Introduce the matrix $M_{\hat{\alpha},\theta_1,\theta_2,\beta}$ given by

$$M_{\hat{\alpha},\theta_1,\theta_2,\beta} := \begin{bmatrix} \mathbf{I} & \mathbf{0} & -A_{\hat{\alpha},\theta_1} \\ \mathbf{0} & \mathbf{I} & -A_{\hat{\alpha},\theta_2} R_{\theta_2-\theta_1-\beta} \\ \mathcal{R}_1 & \mathcal{R}_2 & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{6\times 6} \ , \tag{3.21}$$

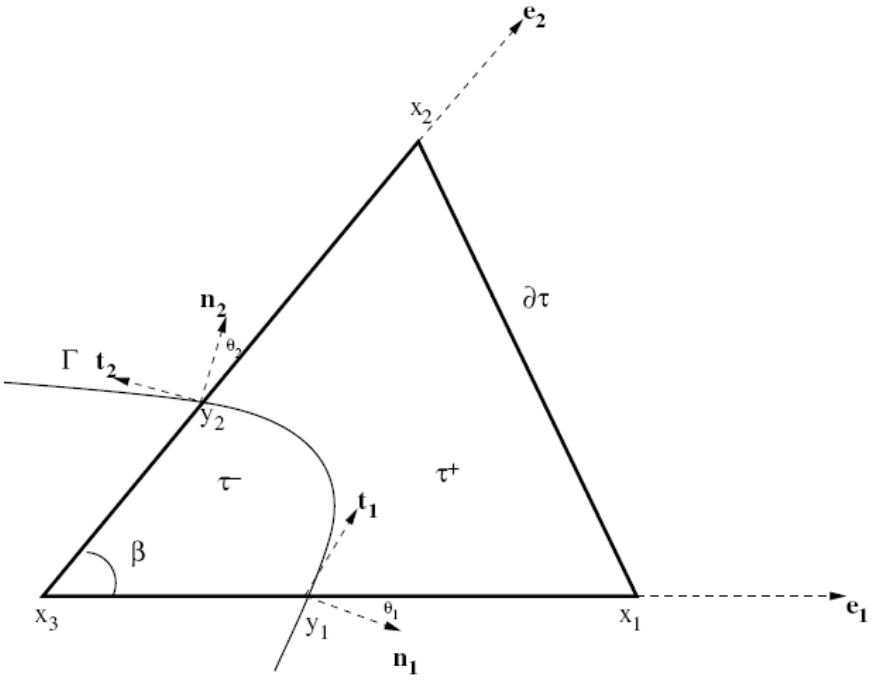


Figure 3-1: *A typical element satisfying Assumption 3.5.*

where $\mathbf{I}$ and $\mathbf{0}$ are the two dimensional identity and zero matrices respectively. The matrices $A_{\hat{\alpha},\theta}$, $\mathcal{R}_1$ and $\mathcal{R}_2$ are given by

$$A_{\hat{\alpha},\theta} := \begin{bmatrix} \cos\theta & \sin\theta \\ \hat{\alpha}\cos\theta & \sin\theta \end{bmatrix}, \quad \mathcal{R}_1 := \begin{bmatrix} r_1^- & r_1^+ \\ 0 & 0 \end{bmatrix}, \qquad \mathcal{R}_2 := \begin{bmatrix} 0 & 0 \\ r_2^- & r_2^+ \end{bmatrix}, \quad (3.22)$$

and R_ϕ is the rotation matrix

$$R_\phi := \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}.$$

Also define the vector $\mathbf{c}(v) := [0,0,0,0,v(x_1^\tau) - v(x_3^\tau), v(x_2^\tau) - v(x_3^\tau)]^T$. The local boundary conditions $\phi_{p,\tau}$ are defined as linear functions on the line segments $\overline{x_3^\tau y_1}$, $\overline{y_1 x_1^\tau}$, $\overline{x_3^\tau y_2}$, $\overline{y_2 x_2^\tau}$, $\overline{x_1^\tau x_2^\tau}$ such that $\phi_{p,\tau}(x_q^\tau) = \delta_{pq}$ and $\sum_{p=1}^3 \phi_{p,\tau} = 1$ on $\partial\tau$. Therefore we can uniquely define $\phi_{p,\tau}$ by the gradients of the linear functions on each line segment, which are found by the following algorithm.

Algorithm 3.6. *Let $\tau \in \mathcal{T}_H^C(\Omega)$ under Assumption 3.5, define the local boundary conditions $\phi_{p,\tau}$ for $p = 1,2,3$ by the following procedure.*

1. *Solve the linear system:*

$$M_{\hat{\alpha},\theta_1,\theta_2,\beta}\mathbf{d}_p = \mathbf{c}(\phi_{p,\tau}) \ . \tag{3.23}$$

2. Then set

$$\begin{cases} (D_{e_1}\phi_{p,\tau})|_{\overline{x_3^\tau y_1}} = (\mathbf{d}_p)_1 \,, & (D_{e_1}\phi_{p,\tau})|_{\overline{y_1 x_1^\tau}} = (\mathbf{d}_p)_2 \,, \\ (D_{e_2}\phi_{p,\tau})|_{\overline{x_3^\tau y_2}} = (\mathbf{d}_p)_3 \,, & (D_{e_2}\phi_{p,\tau})|_{\overline{y_2 x_2^\tau}} = (\mathbf{d}_p)_4 \,, \end{cases} \tag{3.24}$$

where D_e indicates the directional derivative with respect to $\mathbf{e}$. Also let $\phi_{p,\tau}$ be linear between x_1^τ and x_2^τ such that $\phi_{p,\tau}(x_1^\tau) = \delta_{p1}$ and $\phi_{p,\tau}(x_2^\tau) = \delta_{p2}$.

An example of the resulting boundary conditions is shown in Figure 3-2. To see how this algorithm arises, how the matrix $M_{\hat{\alpha},\theta_1,\theta_2,\beta}$ is invertible and how to obtain a robust finite element error estimate where (3.14) holds we examine some of the properties of the exact solution to the interface problem (3.1).

3.1.3 Properties of the exact solution to the interface problem

From now on we denote the restriction of u on $\tau^\pm$ by $u^\pm$ where u is the exact solution to Problem 2.2. The solution to the interface problem (3.1) satisfies the following interface conditions.

Proposition 3.7. *Let u be the exact solution of Problem 2.2. Then u satisfies the interface conditions:*

$$\left(D_n u^+\right)(x) = \alpha^- \left(D_n u^-\right)(x) \quad \text{and} \quad \left(D_t u^+\right)(x) = \left(D_t u^-\right)(x) \tag{3.25}$$

for any $x \in \Gamma$ where $\mathbf{n}$ is the outward normal from Ω^- and $\mathbf{t}$ the corresponding tangent.

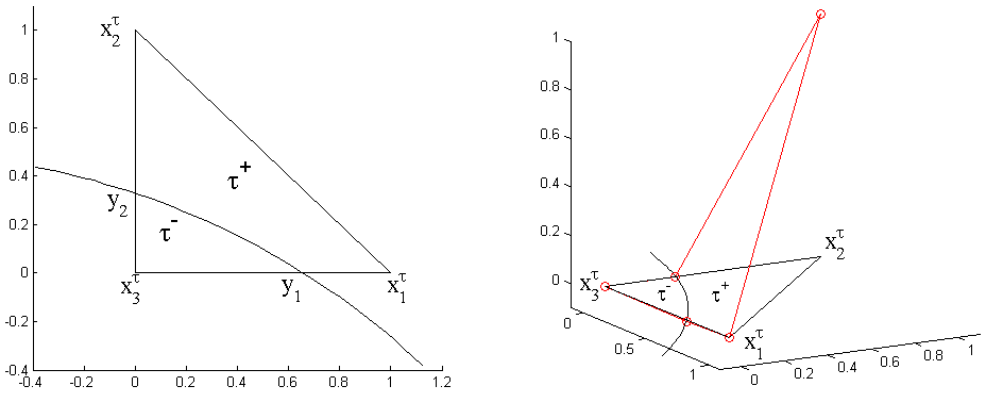
Proof. These conditions can be found in Dautray and Lions (p584 [31]). The tangential derivative condition is more commonly stated simply as a zero jump in u along Γ, $[u]_\Gamma = 0$ from which the tangential derivatives are found. □

The first Lemma allows us to write the edge derivatives of u^- and u^+ at y_i in terms of the normal and tangential derivatives of u at y_i.

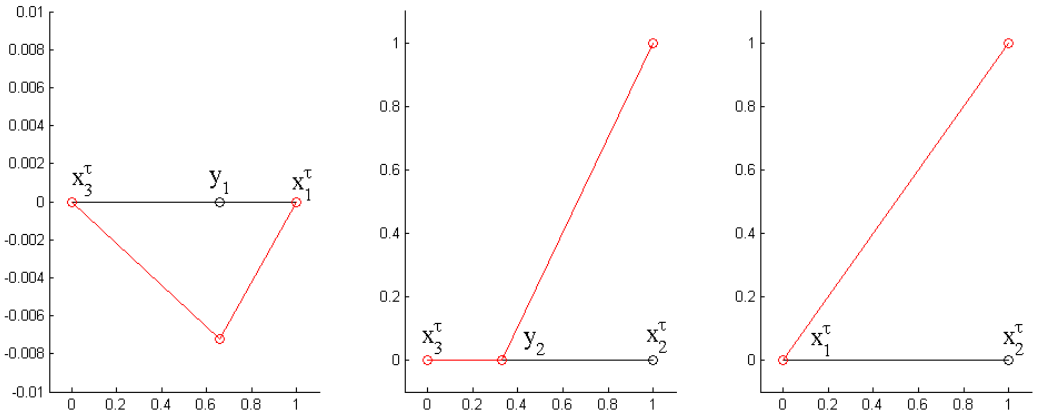
Lemma 3.8. *Let u be the exact solution of Problem 2.2. For $i = 1, 2$,*

$$\begin{bmatrix} D_{e_i}u^-(y_i) \\ D_{e_i}u^+(y_i) \end{bmatrix} = A_{\hat{\alpha},\theta_i} \begin{bmatrix} D_{n_i}u^-(y_i) \\ D_{t_i}u^-(y_i) \end{bmatrix} \tag{3.26}$$

where $A_{\hat{\alpha},\theta}$ is given in (3.22).



(a) XY-plane showing the element and interface (left) and the basis function in 3D (right)



(b) 1D graphs of the basis function along the edges given in the ($x_3^\tau x_1^\tau$-Z) plane (left), ($x_3^\tau x_2^\tau$-Z) plane (middle) and ($x_1^\tau x_2^\tau$-Z) plane (right).

Figure 3-2: *An example of the boundary conditions created by Algorithm 3.6 for $p = 2$ given as 1D graphs along each edge (the element edges are given in black and the function values in red). For this example $\hat{\alpha} = 100$.*

Proof. The proof follows from noting that $D_{e_i}(\cdot) = \cos\theta_i D_{n_i}(\cdot) + \sin\theta_i D_{t_i}(\cdot)$ from (3.20) and combining this with the interface conditions (3.25). □

Using this lemma we can start to see how the linear system (3.23) arises. For any $v \in H_0^1(\Omega)$ with suitably defined point values at y_i define

$$\mathbf{d}(v) := \left[D_{e_1}v^-(y_1),\ D_{e_1}v^+(y_1),\ D_{e_2}v^-(y_2),\ D_{e_2}v^+(y_2),\ D_{n_1}v^-(y_1),\ D_{t_1}v^-(y_1)\right]^T \tag{3.27}$$

By Lemma 3.8 for $i = 1$ we have

$$\left[\begin{array}{ccc} \mathbf{I} & \mathbf{0} & -A_{\hat{\alpha},\theta_1} \end{array} \right] \mathbf{d}(u) = \left[\begin{array}{c} 0 \\ 0 \end{array} \right] , \tag{3.28}$$

which if we replace $\mathbf{d}(u)$ by $\sum_{p=1}^{3} u(x_p^\tau)\mathbf{d}_p$ leads to the first two equations in (3.23). The next property allows us to relate the directional derivatives $D_{n_2}u(y_2)$ and $D_{t_2}u(y_2)$ to the derivatives $D_{n_1}u(y_1)$ and $D_{t_1}u(y_1)$ by a simple application of Taylor's theorem.

Lemma 3.9. *Let u be the exact solution of Problem 2.2. Then*

$$\left[\begin{array}{c} D_{n_2}u^-(y_2) \\ D_{t_2}u^-(y_2) \end{array} \right] = R_{\theta_2-\theta_1-\beta} \left[\begin{array}{c} D_{n_1}u^-(y_1) \\ D_{t_1}u^-(y_1) \end{array} \right] + \epsilon' , \tag{3.29}$$

where

$$\begin{aligned} \left\|\epsilon'\right\|_\infty \lesssim H_\tau^{\frac{1}{2}} \Big[& \left\|D_{e_2}D_{n_1}u^-\right\|^2_{L_2(e_2\cap\tau^-)} + \left\|D_{e_1}D_{n_1}u^-\right\|^2_{L_2(e_1\cap\tau^-)} \\ & + \left\|D_{e_2}D_{t_1}u^-\right\|^2_{L_2(e_2\cap\tau^-)} + \left\|D_{e_1}D_{t_1}u^-\right\|^2_{L_2(e_1\cap\tau^-)} \Big]^{\frac{1}{2}} . \end{aligned}$$

Proof. Note that $R_\phi \mathbf{n}_i = \cos\phi\mathbf{n}_i + \sin\phi\mathbf{t}_i$ and $R_\phi\mathbf{t}_i = -\sin\phi\mathbf{n}_i + \cos\phi\mathbf{t}_i$ for $i = 1, 2$. Then $\mathbf{n}_2 = R_{-\theta_2}\mathbf{e}_2 = R_{-\theta_2+\beta}\mathbf{e}_1 = R_{-\theta_2+\beta+\theta_1}\mathbf{n}_1$ and this implies

$$\begin{aligned} \mathbf{n}_2 &= \cos(\theta_2 - \theta_1 - \beta)\mathbf{n}_1 - \sin(\theta_2 - \theta_1 - \beta)\mathbf{t}_1 , \\ \mathbf{t}_2 &= \sin(\theta_2 - \theta_1 - \beta)\mathbf{n}_1 + \cos(\theta_2 - \theta_1 - \beta)\mathbf{t}_1 , \end{aligned}$$

where we used $\sin(-\phi) = -\sin(\phi)$ and $\cos(-\phi) = \cos(\phi)$. This gives

$$\left[\begin{array}{c} D_{n_2}u^-(x) \\ D_{t_2}u^-(x) \end{array} \right] = R_{\theta_2-\theta_1-\beta} \left[\begin{array}{c} D_{n_1}u^-(x) \\ D_{t_1}u^-(x) \end{array} \right] \tag{3.30}$$

for any $x \in \tau^-$. The result then follows by letting $x = y_2$ and using Taylor expansions on the right hand side of (3.30) along e_2 about x_3^τ to obtain

$$\left|D_{n_1}u^-(y_2)\right| \lesssim \left|D_{n_1}u^-(x_3^\tau)\right| + H_\tau^{\frac{1}{2}} \left\|D_{e_2}D_{n_1}u^-\right\|_{L_2(e_2\cap\tau^-)} ,$$

and then along e_1 about y_1 to obtain

$$\left|D_{n_1}u^-(y_2)\right| \lesssim \left|D_{n_1}u^-(y_1)\right| + H_\tau^{\frac{1}{2}} \left\|D_{e_1}D_{n_1}u^-\right\|_{L_2(e_1\cap\tau^-)} + H_\tau^{\frac{1}{2}} \left\|D_{e_2}D_{n_1}u^-\right\|_{L_2(e_2\cap\tau^-)} .$$

An analogous result is obtained for $|D_{t_1}u^-(y_2)|$. The Taylor expansion is possible because u^- is in H^2 on each $e_i \cap \tau^-$. Finally since $|R_{\theta_2-\theta_1-\beta}|_\infty \leq 1$ the obtain the

required bound for ϵ'. □

Using Lemmas 3.8 and 3.9 with $i = 2$ we can start to see where the third and forth rows of (3.23) come from, by noting that

$$\begin{bmatrix} \mathbf{0} & \mathbf{I} & -A_{\hat{\alpha},\theta_2} R_{\theta_2-\theta_1-\beta} \end{bmatrix} \mathbf{d}(u) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + A_{\hat{\alpha},\theta_2} \epsilon' \ . \tag{3.31}$$

If we then set $\epsilon' = 0$ then the solutions of the third and forth rows of (3.23) satisfy (3.31) when $\mathbf{d}(u)$ is replaced by $\sum_{p=1}^{3} u(x_p^\tau)\mathbf{d}_p$. The final two equations come from the following lemma.

Lemma 3.10. *Let u be the exact solution of Problem 2.2 and define $\epsilon \in \mathbb{R}^2$ such that*

$$r_i^- \left(D_{e_i} u^-\right)(y_i) + r_i^+ \left(D_{e_i} u^+\right)(y_i) = u(x_i^\tau) - u(x_3^\tau) + \epsilon_i \tag{3.32}$$

for $i = 1, 2$. Then

$$|\epsilon_i| \lesssim H_\tau^{\frac{3}{2}} \left(\left\| D_{e_i}^2 u^- \right\|_{L_2(e_i \cap \tau^-)} + \left\| D_{e_i}^2 u^+ \right\|_{L_2(e_i \cap \tau^+)} \right) \tag{3.33}$$

for $i = 1, 2$.

Proof. The result follows from Taylor expansions at the point y_i on each side of the interface and the interface matching condition $u^+(y_i) = u^-(y_i)$. The remainder follows from the fact that $u^\pm \in H^2(\tau^\pm)$. □

This produces the matrix system

$$\begin{bmatrix} \mathcal{R}_1 & \mathcal{R}_2 & \mathbf{0} \end{bmatrix} \mathbf{d}(u) = \begin{bmatrix} u(x_1^\tau) - u(x_3^\tau) \\ u(x_2^\tau) - u(x_3^\tau) \end{bmatrix} + \epsilon \ . \tag{3.34}$$

Combining (3.28), (3.31) and (3.34) we get the following corollary.

Corollary 3.11. *Let u be the exact solution of Problem 2.2, then for each cut element $\tau \in \mathcal{T}_H^C(\Omega)$ that satisfies Assumption 3.5 we have*

$$M_{\hat{\alpha},\theta_1,\theta_2,\beta} \mathbf{d}(u) = \mathbf{c}(u) + \boldsymbol{\delta} \ , \tag{3.35}$$

where $\boldsymbol{\delta} \in \mathbb{R}^6$ is defined by

$$\boldsymbol{\delta} = \begin{bmatrix} \mathbf{0} \\ A_{\hat{\alpha},\theta_2} \epsilon' \\ \epsilon \end{bmatrix} \ . \tag{3.36}$$

Using the linear system (3.23), Corollary 3.11 and that $\mathbf{c}(u)$ depends only on the nodal values of u we have

$$M_{\hat{\alpha},\theta_1,\theta_2,\beta}\left(\mathbf{d}(u)-\sum_{p=1}^{3}u(x_p^\tau)\mathbf{d}_p\right)=\mathbf{c}(u)+\boldsymbol{\delta}-\mathbf{c}(I_H^{\mathrm{MS}}u)=\mathbf{c}(u)+\boldsymbol{\delta}-\mathbf{c}(u)=\delta\ .$$

Thus the boundary error can be analysed by considering $M^{-1}_{\hat{\alpha},\theta_1,\theta_2,\beta}\boldsymbol{\delta}$.

3.1.4 Boundary error for the artificial local boundary conditions

In the previous section we have seen how the artificial boundary condition for the subgrid problem is constructed. It requires a rudimentary linear solve of a six by six matrix system to provide the gradients of $\phi_{p,\tau}$ on $\overline{x_3^\tau y_1}$, $\overline{y_1 x_1^\tau}$, $\overline{x_3^\tau y_2}$, $\overline{y_2 x_2^\tau}$, $\overline{x_1^\tau x_2^\tau}$. Now we begin to explore the harder aspect of [27], the analysis that shows this is a good choice of boundary condition to achieve an accurate approximation. In this section we explore the solvability of the linear system (3.23) and a resulting bound for the error on the boundary of $E_H^{\mathrm{MS}}=u-I_H^{\mathrm{MS}}u$ (recall Remark 3.4). The first result we explore is Theorem 3.6 from [27]. It establishes the solvability of these systems and provides a bound on the solution.

Theorem 3.12. *Under Assumption 3.5, suppose $\phi:=\theta_2-\theta_1-\beta\neq 0$ and introduce the 2×2 matrix*

$$D:=\mathcal{R}_1A_{\hat{\alpha},\theta_1}+\mathcal{R}_2A_{\hat{\alpha},\theta_2}R_\phi\ .$$

Then, for all sufficiently large $\hat{\alpha}$, both D and $M_{\hat{\alpha},\theta_1,\theta_2,\beta}$ are nonsingular with

$$(M_{\hat{\alpha},\theta_1,\theta_2,\beta})^{-1}=\begin{bmatrix}\mathbf{I}&\mathbf{0}&A_{\hat{\alpha},\theta_1}\\ \mathbf{0}&\mathbf{I}&A_{\hat{\alpha},\theta_2}R_\phi\\ \mathbf{0}&\mathbf{0}&\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{I}&\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{I}&\mathbf{0}\\ -D^{-1}\mathcal{R}_1&-D^{-1}\mathcal{R}_2&D^{-1}\end{bmatrix}\ . \tag{3.37}$$

Moreover

$$\left\|D^{-1}\right\|_\infty\lesssim\frac{1}{\hat{\alpha}}\frac{1}{H_\tau}\left(\sin(\phi)\right)^{-1}\ . \tag{3.38}$$

Proof. An elementary calculation shows that for

$$E:=\begin{bmatrix}r_1^+\cos\theta_1&0\\ r_2^+\cos\theta_2\cos\phi&-r_2^+\cos\theta_2\sin\phi\end{bmatrix}\ ,$$

we have $\left\|\hat{\alpha}^{-1}D - E\right\|_\infty \leq C\hat{\alpha}^{-1}H_\tau$ where C is independent of $\theta_1, \theta_2, \phi, \beta$ and H_τ. Matrix perturbation theory shows that the set of invertible operators is open. Combining this along with the contraction mapping theorem we see that D is invertible if and only if E is invertible and $\left\|\hat{\alpha}^{-1}D - E\right\|_\infty \leq \left\|E^{-1}\right\|_\infty^{-1}$. Hence

$$E^{-1} := \begin{bmatrix} \left(r_1^+ \cos\theta_1\right)^{-1} & 0 \\ -\left(r_1^+ \cos\theta_1 \sin\phi\right)^{-1} \cos\phi & -\left(r_2^+ \cos\theta_2 \sin\phi\right)^{-1} \end{bmatrix},$$

and so $\left\|E^{-1}\right\|_\infty \leq H_\tau^{-1} (\sin\phi)^{-1}$. This imples $\left\|\hat{\alpha}^{-1}D - E\right\|_\infty \leq CH_\tau \leq \left\|E^{-1}\right\|_\infty^{-1}$ for sufficiently large $\hat{\alpha}$. Then

$$\left\|\hat{\alpha}D^{-1}\right\|_\infty = \left\|\left(\hat{\alpha}^{-1}D\right)^{-1}\right\|_\infty \leq C\left\|E^{-1}\right\|_\infty \lesssim H_\tau^{-1}(\sin\phi)^{-1}$$

using the upper and lower bounds introduced in Assumption 3.5. Since D^{-1} exists then the formula for $\left(M_{\hat{\alpha},\theta_1,\theta_2,\beta}\right)^{-1}$ is verified by simple matrix multiplication. □

Chu, Graham and Hou make several remarks about Algorithm 3.6 and certain special cases that are worth noting.

Remark 3.13. *1. If $\theta_i = 0$ for $i = 1, 2$ the normal coincides with the edge e_i. In this case the boundary condition computed from Algorithm 3.6 coincides with the "oscillatory boundary condition" given in [49] that were introduced in Section 1.2.2. So if $\theta_1 = 0$ then the first two and last two equations of 3.23 imply*

$$(\mathbf{d}_p)_2 = \hat{\alpha}\,(\mathbf{d}_p)_1 \qquad \text{and} \qquad r_1^-\,(\mathbf{d}_p)_1 + r_1^+\,(\mathbf{d}_p)_2 = \phi_{p,\tau}(x_1) - \phi_{p,\tau}(x_3)$$

and so

$$(\mathbf{d}_p)_1 = \frac{\phi_{p,\tau}(x_1) - \phi_{p,\tau}(x_3)}{r_1^- + \hat{\alpha}r_1^+}, \qquad (\mathbf{d}_p)_2 = \hat{\alpha}\left(\frac{\phi_{p,\tau}(x_1) - \phi_{p,\tau}(x_3)}{r_1^- + \hat{\alpha}r_1^+}\right).$$

Thus $\phi_{p,\tau}$ is the solution of $-\left(\alpha\phi'_{p,\tau}\right)' = 0$ on $\overline{x_3^\tau x_1^\tau}$.

2. When $\theta_i \neq 0$ for both $i = 1, 2$ then the boundary condition on each $\mathbf{e}_i$ depends on both θ_1 and θ_2. This is because of Lemma 3.9 which links the boundary conditions on each edge to the normal derivatives on $\mathbf{e}_1$. This suggests that a purely local boundary condition that samples α only on isolated edges may not be sufficient to generate a robust error estimate.

3. Algorithm 3.6 determines the boundary conditions $\phi_{p,\tau}$ on each element $\tau \in \mathcal{T}_H(\Omega)$ separately and so the resulting basis functions Φ_p^{MS} may not be contin-

uous across element edges. Therefore approximation in the space spanned by these functions may not be conforming. This problem is solved by averaging the boundary conditions between neighbouring elements thus producing conforming finite elements. We discuss this further in Section 3.1.6.

The critical case in Theorem 3.12, when $\theta_2 - \theta_1 - \beta = 0$, occurs when the unit outward normals $\mathbf{n}_1$ and $\mathbf{n}_2$ at the two intersection points y_1 and y_2 are in the same direction. They remark that in the case where the interface is not a straight line then the mesh may be refined such that the element is subdivided into two sub-elements where $\theta_2 - \theta_1 - \beta \neq 0$ in each sub-element.

However if the interface is a straight line through an element then we have to take the approach of resolving the interface with the mesh by subdividing the quadrilateral created in τ^+ into two triangles, one whos edge is the interface and use this along with τ^- as a refinement of the mesh.

In [[27] Remark 3.13] it is mentioned that the question of the (non)singularity of the matrix $M_{\hat{\alpha},\theta_1,\theta_2,\beta}$ has not been analysed under the general assumption that only $\phi = \theta_2 - \theta_1 - \beta = 0$ for general choices of $\hat{\alpha}$, θ_i, r_i^- and r_i^+. We explore this in the following lemma.

Lemma 3.14. *Under Assumption 3.5, suppose $\phi := \theta_2 - \theta_1 - \beta = 0$ then the matrix $M_{\hat{\alpha},\theta_1,\theta_2,\beta}$ is invertible if for $r_i = r_i^- + r_i^+$, $i = 1, 2$,*

$$\frac{r_1}{r_1^- + \hat{\alpha} r_1^+} \neq \frac{r_2}{r_2^- + \hat{\alpha} r_2^+} \ .$$

Proof. From Theorem 3.12 we know the the matrix $M_{\hat{\alpha},\theta_1,\theta_2,\beta}$ is invertible if the 2×2 matrix D is itself invertible. From the proof of Theorem 3.12 we have that

$$\begin{aligned} D = \mathcal{R}_1 A_{\hat{\alpha},\theta_1} + \mathcal{R}_2 A_{\hat{\alpha},\theta_2} R_\phi &= \mathcal{R}_1 A_{\hat{\alpha},\theta_1} + \mathcal{R}_2 A_{\hat{\alpha},\theta_2} \\ &= \begin{bmatrix} \left(r_1^- + \hat{\alpha} r_1^+\right) \cos\theta_1 & r_1 \sin\theta_1 \\ \left(r_2^- + \hat{\alpha} r_2^+\right) \cos\theta_2 & r_2 \sin\theta_2 \end{bmatrix} . \end{aligned}$$

Therefore the determinant of D is

$$\det(D) = \left(r_1^- + \hat{\alpha} r_1^+\right) r_2 \cos\theta_1 \sin\theta_2 - r_1 \left(r_2^- + \hat{\alpha} r_2^+\right) \sin\theta_1 \cos\theta_2 \ ,$$

and this is zero only if

$$\left(\frac{r_1}{r_1^- + \hat{\alpha} r_1^+}\right) \tan\theta_1 = \left(\frac{r_2}{r_2^- + \hat{\alpha} r_2^+}\right) \tan\theta_2 \ .$$

Now this is always equal if $\theta_1 = \theta_2 = 0$ but as we have seen in Remark 3.13 this reduces to the case of "oscillatory boundary conditions". Therefore the matrix D is singular only if

$$\frac{r_1}{r_1^- + \hat{\alpha} r_1^+} = \frac{r_2}{r_2^- + \hat{\alpha} r_2^+}$$

and has a solution otherwise. □

Remark 3.15. *The above lemma shows that when $\phi = \theta_2 - \theta_1 - \beta = 0$ and $\hat{\alpha}$ is very large, a sufficient condition for invertibility of $M_{\hat{\alpha},\theta_1,\theta_2,\beta}$ is that the two ratios r_1/r_1^+ and r_2/r_2^+ are not approximately equal. i.e. D is ill conditioned when*

$$\frac{r_1}{r_1^+} \approx \frac{r_2}{r_2^+} \ .$$

We also note that in the case when $\phi = 0$ we do not know if $\|D\|_\infty = O(\hat{\alpha}^{-1})$ and so in this case it is not possible to prove a robust error estimate even though the matrix system (3.23) is invertible. When using MsFEM it is recommended to avoid the case $\phi = 0$ by perturbing the node x_3^τ or if the interface is a straight line then refine the mesh to resolve the interface.

The next theorem shows that the nodal interpolant $I_H^{\mathrm{MS}} u$ from Remark 3.4 is a good approximation to u along the boundary of the element τ. From now on we exclusively refer to $E_H^{\mathrm{MS}} := u - I_H^{\mathrm{MS}} u$.

Theorem 3.16. *Let u be the exact solution of Problem 2.2. Suppose an element τ intersects the interface as in Assumption 3.5 and suppose $\phi \neq 0$. Then we have for $m = 0, 1$*

$$\begin{aligned} &\max_{i=1,2} H_\tau^m \left\{ \hat{\alpha} \left\| D_{e_i}^m E_H^{MS} \right\|_{L_\infty(e_i \cap \tau^-)} \ , \ \left\| D_{e_i}^m E_H^{MS} \right\|_{L_\infty(e_i \cap \tau^+)} \right\} \\ &\quad \lesssim H_\tau^{3/2} \max_{\substack{i=1,2 \\ |k|=1}} \left[\hat{\alpha}^2 \left\| D^k D_{e_i} u \right\|^2_{L_2(e_i \cap \tau^-)} + \left\| D^k D_{e_i} u \right\|^2_{L_2(e_i \cap \tau^+)} \right]^{\frac{1}{2}} . \end{aligned} \tag{3.39}$$

Proof. This theorem essentially shows the derivative of the error $D_{e_i} E_H^{\mathrm{MS}}$ is still of optimal order, $O(H_\tau^{\frac{1}{2}})$, on the edges of τ despite the fact that u has jumping derivatives along each edge $\mathbf{e}_i$. It also shows the explicit dependence on $\hat{\alpha}$ only on τ^-. A robust error estimate on the edges is then found by showing that $\left\| D^k D_{e_i} u \right\|^2_{L_2(e_i \cap \tau^-)} = O(\hat{\alpha}^{-1})$ (shown later in Theorem 3.22 using Corollary 3.17).

The proof is given on the assumption that $\tau^\pm$ are as in Figure 3-1 where α is large in the region containing x_3. Using the linear system (3.23), Corollary 3.11 and that $\mathbf{c}(u)$

depends only on the nodal values of u we have

$$M_{\hat{\alpha},\theta_1,\theta_2,\beta}\left(\mathbf{d}(u)-\sum_{p=1}^{3}u(x_p^\tau)\mathbf{d}_p\right)=\mathbf{c}(u)+\boldsymbol{\delta}-\mathbf{c}(I_H^{\mathrm{MS}}u)=\mathbf{c}(u)+\boldsymbol{\delta}-\mathbf{c}(u)=\boldsymbol{\delta}\ .$$

Hence by Theorem 3.12 we invert the above system to obtain

$$\left(\mathbf{d}(u)-\sum_{p=1}^{3}u(x_p^\tau)\mathbf{d}_p\right)=\begin{bmatrix}\mathbf{I}&\mathbf{0}&A_{\hat{\alpha},\theta_1}\\ \mathbf{0}&\mathbf{I}&A_{\hat{\alpha},\theta_2}R_\phi\\ \mathbf{0}&\mathbf{0}&\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ A_{\hat{\alpha},\theta_2}\epsilon'\\ D^{-1}(\epsilon-\mathcal{R}_2A_{\hat{\alpha},\theta_2}\epsilon')\end{bmatrix}\ . \tag{3.40}$$

Expanding the left hand side of (3.40), using (3.23), (3.24) and (3.27), we have that the first four entries are $D_{e_1}(u-I_H^{\mathrm{MS}}u)^-(y_1)$, $D_{e_1}(u-I_H^{\mathrm{MS}}u)^+(y_1)$, $D_{e_2}(u-I_H^{\mathrm{MS}}u)^-(y_2)$ and $D_{e_2}(u-I_H^{\mathrm{MS}}u)^+(y_2)$. Expanding the right hand side of (3.40) gives the first two entries as

$$A_{\hat{\alpha},\theta_1}D^{-1}(\epsilon-\mathcal{R}_2A_{\hat{\alpha},\theta_2}\epsilon')\ .$$

Recalling Lemmas 3.9, 3.10 and (3.38) we have that

$$\begin{aligned}\left\|D^{-1}(\epsilon-\mathcal{R}_2A_{\hat{\alpha},\theta_2}\epsilon')\right\|_\infty&\le\left\|D^{-1}\right\|_\infty\left\|\epsilon-\mathcal{R}_2A_{\hat{\alpha},\theta_2}\epsilon'\right\|_\infty\\&\le\left\|D^{-1}\right\|_\infty\left(\|\epsilon\|_\infty+\left\|\mathcal{R}_2A_{\hat{\alpha},\theta_2}\epsilon'\right\|_\infty\right)\ ,\end{aligned}\tag{3.41}$$

where (3.38) implies

$$\left\|D^{-1}\right\|_\infty\lesssim\hat{\alpha}^{-1}H_\tau^{-1}\ . \tag{3.42}$$

Lemma 3.10 implies

$$\|\epsilon\|_\infty\lesssim H_\tau^{\frac{3}{2}}\left\{\left\|D_{e_i}^2u^-\right\|_{L_2(e_i\cap\tau^-)}+\left\|D_{e_i}^2u^+\right\|_{L_2(e_i\cap\tau^+)}\right\}\ . \tag{3.43}$$

Finally using (3.22) we obtain

$$\begin{aligned}\left\|\mathcal{R}_2A_{\hat{\alpha},\theta_2}\epsilon'\right\|_\infty&\lesssim\|\mathcal{R}_2\|_\infty\|A_{\hat{\alpha},\theta_2}\|_\infty\left\|\epsilon'\right\|_\infty\\&\lesssim H_\tau\hat{\alpha}\left\|\epsilon'\right\|_\infty\\&\lesssim H_\tau^{\frac{3}{2}}\hat{\alpha}\left\{\left\|D_{e_2}D_{n_1}u^-\right\|^2_{L_2(e_2\cap\tau^-)}+\left\|D_{e_1}D_{n_1}u^-\right\|^2_{L_2(e_1\cap\tau^-)}\right.\\&\qquad\left.+\left\|D_{e_2}D_{t_1}u^-\right\|^2_{L_2(e_2\cap\tau^-)}+\left\|D_{e_1}D_{t_1}u^-\right\|^2_{L_2(e_1\cap\tau^-)}\right\}^{\frac{1}{2}}\end{aligned}\tag{3.44}$$

by Lemma 3.9. Substituting (3.42), (3.43) and (3.44) back into (3.41) we obtain

$$\left\|D^{-1}(\epsilon - \mathcal{R}_2 A_{\hat{\alpha},\theta_2}\epsilon')\right\|_\infty \lesssim H_\tau^{\frac{1}{2}} \max_{\substack{i=1,2\\|k|=1}} \left[\left\|D^k D_{e_i} u\right\|^2_{L_\infty(e_i\cap\tau^-)} + \hat{\alpha}^{-2}\left\|D^k D_{e_i} u\right\|^2_{L_\infty(e_i\cap\tau^+)}\right]^{\frac{1}{2}} .$$

Note that in the previous estimate we have used the commutativity of directional derivatives along with the fact that $|D_e v| \leq |\partial v/\partial x| + |\partial v/\partial y| \leq 2\max_{|k|=1}\left|D^k v\right|$ for any normalised vector $\mathbf{e}$. Specifically we substitute $\mathbf{e}_i$, $\mathbf{n}_1$ and $\mathbf{t}_1$ into this inequality. Hence, again using (3.22),

$$\max\left\{\hat{\alpha}\left|D_{e_1}(u - I_H^{\mathrm{MS}}u)^-(y_1)\right|,\ \left|D_{e_1}(u - I_H^{\mathrm{MS}}u)^+(y_1)\right|\right\} \lesssim H_\tau^{\frac{1}{2}} \max_{\substack{i=1,2\\|k|=1}} \left[\hat{\alpha}^2\left\|D^k D_{e_i} u\right\|^2_{L_\infty(e_i\cap\tau^-)} + \left\|D^k D_{e_i} u\right\|^2_{L_\infty(e_i\cap\tau^+)}\right]^{\frac{1}{2}} .$$

Similarly expanding the third and forth terms of the right hand side of (3.40) we obtain

$$A_{\hat{\alpha},\theta_2}\epsilon' + A_{\hat{\alpha},\theta_2} R_\phi D^{-1}(\epsilon - \mathcal{R}_2 A_{\hat{\alpha},\theta_2}\epsilon') ,$$

which by a similar expansion and manipulation gives the same bound:

$$\max\left\{\hat{\alpha}\left|D_{e_2}(u - I_H^{\mathrm{MS}}u)^-(y_2)\right|,\ \left|D_{e_2}(u - I_H^{\mathrm{MS}}u)^+(y_2)\right|\right\} \lesssim H_\tau^{\frac{1}{2}} \max_{\substack{i=1,2\\|k|=1}} \left[\hat{\alpha}^2\left\|D^k D_{e_i} u\right\|^2_{L_\infty(e_i\cap\tau^-)} + \left\|D^k D_{e_i} u\right\|^2_{L_\infty(e_i\cap\tau^+)}\right]^{\frac{1}{2}} .$$

Then (3.39) for $m = 1$ follows by a simple application of the fundamental theorem of calculus. For example if $x \in e_i \cap \tau^-$ we have

$$\begin{aligned} D_{e_i}(u - I_H^{\mathrm{MS}}u)^-(x) &= D_{e_i}(u - I_H^{\mathrm{MS}}u)^-(y_i) - \int_x^{y_i} D_{e_i}^2(u - I_H^{\mathrm{MS}}u)^-(z)\ dz \\ &= D_{e_i}(u - I_H^{\mathrm{MS}}u)^-(y_i) - \int_x^{y_i} D_{e_i}^2 u^-(z)\ dz \end{aligned}$$

since $I_H^{\mathrm{MS}}u$ is linear on $e_i \cap \tau^-$. Then

$$\begin{aligned} \hat{\alpha}\left|D_{e_i}(u - I_H^{\mathrm{MS}}u)^-(x)\right| &\lesssim \hat{\alpha}\left|D_{e_i}(u - I_H^{\mathrm{MS}}u)^-(y_i)\right| + \hat{\alpha}\left\|D_{e_i}^2 u\right\|_{L_2(e_i\cap\tau^-)}\left(\int_x^{y_i} dz\right)^{\frac{1}{2}} \\ &\lesssim \hat{\alpha}\left|D_{e_i}(u - I_H^{\mathrm{MS}}u)^-(y_i)\right| + H_\tau^{\frac{1}{2}}\hat{\alpha}\left\|D^k D_{e_i} u\right\|_{L_2(e_i\cap\tau^-)} \\ &\lesssim H_\tau^{1/2} \max_{\substack{i=1,2\\|k|=1}} \left[\hat{\alpha}^2\left\|D^k D_{e_i} u\right\|^2_{L_2(e_i\cap\tau^-)} + \left\|D^k D_{e_i} u\right\|^2_{L_2(e_i\cap\tau^+)}\right]^{\frac{1}{2}} . \end{aligned}$$

The results on $e_i \cap \tau^+$ follow in a similar way. To obtain the estimates for $m = 0$ we again use the fundamental theorem of calculus and the fact that $E_H^{\mathrm{MS}}(x_3^\tau) = 0$ to write

$$|E_H^{\mathrm{MS}}(x)| = |E_H^{\mathrm{MS}}(x) - E_H^{\mathrm{MS}}(x_3^\tau)| = \left| \int_{x_3^\tau}^{x} D_{e_i}(u - I_H^{\mathrm{MS}} u)^-(z)\, dz \right| \\ \leq \left\| D_{e_i}(u - I_H^{\mathrm{MS}} u)^- \right\|_{L_\infty(e_i \cap \tau^-)} H_\tau \ .$$

The results for $m = 0$ on τ^+ follow in an analogous way. □

Next we show that we can then bound the edge derivatives on the right hand side of (3.39) by Sobolev norms on the interior of τ. The motivation for this comes from the regularity result in Theorem 2.22, which shows the Sobolev norms on each inclusion Ω_i are $O(\alpha_i^{-1})$. Eventually this regularity result will remove the dependence of (3.39) on $\hat{\alpha}$ (see Theorem 3.22).

Corollary 3.17. *Let u be the exact solution of Problem 2.2. Suppose an element τ intersects the interface as in Assumption 3.5 and suppose $\phi \neq 0$. Then we have for $m = 0, 1$*

$$\max_{i=1,2} H_\tau^m \left\{ \hat{\alpha} \left\| D_{e_i}^m E_H^{MS} \right\|_{L_\infty(e_i \cap \tau^-)} \ , \ \left\| D_{e_i}^m E_H^{MS} \right\|_{L_\infty(e_i \cap \tau^+)} \right\} \\ \lesssim H_\tau \left[\hat{\alpha}^2 \left(|u|^2_{H^2(\tau^-)} + H_\tau \, |u|^2_{H^{5/2}(\tau^-)} \right) + \left(|u|^2_{H^2(\tau^+)} + H_\tau \, |u|^2_{H^{5/2}(\tau^+)} \right) \right]^{\frac{1}{2}} \ . \tag{3.45}$$

Proof. The proof consists of bounding $\left\| D^k D_{e_i} u \right\|^2_{L_2(e_i \cap \tau^-)}$ and $\left\| D^k D_{e_i} u \right\|^2_{L_2(e_i \cap \tau^+)}$ in (3.39). Let $\eta^\pm$ be a polygon chosen inside $\tau^\pm$ such that $\partial\tau \cap \tau^\pm \subset \partial\eta^\pm$. These polygons may be chosen such that $|\eta^\pm| \sim |\tau^\pm|$ (see Figure 3-3).

We also recall the trace theorem for polygons after scaling to any element $\tau \in \mathcal{T}_H(\Omega)$ that gives

$$|v|^2_{H^1(e)} \lesssim H_\tau^{-3} \|v\|^2_{L_2(\tau)} + H_\tau^{-1} |v|^2_{H^1(\tau)} + |v|^2_{H^{3/2}(\tau)} \quad , \quad \text{for all } \ v \in H^{3/2}(\tau)$$

for any edge e. Replacing v by $v - \gamma$ where γ is the constant that appears in the Poincaré inequality to give

$$|v|^2_{H^1(e)} \lesssim H_\tau^{-1} |v|^2_{H^1(\tau)} + |v|^2_{H^{3/2}(\tau)} \quad , \quad \text{for all } \ v \in H^{3/2}(\tau) \ .$$

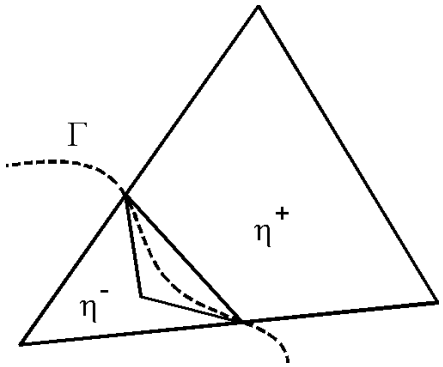


Figure 3-3: *An example of how the polygons η^- and η^+ may be chosen in τ^- and τ^+ respectively.*

Using this estimate along with $|k| = 1$ and $i = 1, 2$ we have

$$\begin{aligned}\left\|D^k D_{e_i} u^-\right\|^2_{L_2(e_i\cap\tau^-)} &= \left|D^k u^-\right|^2_{H^1(e_i\cap\tau^-)} \\ &\lesssim H_\tau^{-1}\left|D^k u^-\right|^2_{H^1(\eta^-)} + \left|D^k u^-\right|^2_{H^{3/2}(\eta^-)} \\ &\lesssim H_\tau^{-1}\left|u^-\right|^2_{H^2(\tau^-)} + \left|u^-\right|^2_{H^{5/2}(\tau^-)}.\end{aligned}$$

Analogously

$$\left\|D^k D_{e_i} u^+\right\|^2_{L_2(e_i\cap\tau^+)} \lesssim H_\tau^{-1}\left|u^+\right|^2_{H^2(\tau^+)} + \left|u^+\right|^2_{H^{5/2}(\tau^+)}.$$

Substituting this into Theorem 3.16 gives the required result. □

3.1.5 Interior error for the artificial local boundary conditions

Since Corollary 3.17 shows that the derivative of $E_H^{\mathrm{MS}} = u - I_H^{\mathrm{MS}} u$ on $\partial\tau$ is robust (after application of Theorem 2.22), what remains now is to show that E_H^{MS} has an extension $\widetilde{E}_H^{\mathrm{MS}}$ (recall Theorem 3.3) that is suitably robust, i.e. such that (3.14) holds. In order to do this we need an additional assumption.

Assumption 3.18. *We impose Assumption 3.5 for $\tau \in \mathcal{T}_H^C(\Omega)$ and assume the interface $\Gamma \cap \tau$ is star-shaped about x_3^τ. So introducing polar coordinates with origin x_3^τ and polar angle θ measured anticlockwise from $\mathbf{e}_1$ we assume that each $(x, y) = (r(\theta)cos\theta, r(\theta)sin\theta)$, for $\theta \in [0, \beta]$. The authors of [27] show that this leads to*

$$r(\theta) \sim H_\tau \qquad \textit{for all} \quad \theta \in [0, \beta].$$

Also letting s denote the arclength along $\Gamma \cap \tau$ they show that

$$ds = \sqrt{(r(\theta))^2 + (r'(\theta))^2}\, d\theta \sim H_\tau d\theta$$

and that this assumption leads to

$$|\Gamma \cap \tau| \sim H_\tau \ , \qquad \left|\tau^{\pm}\right| \sim H_\tau^2 \ .$$

What we present next is a generalisation of Lemma 3.15 in [27]. As presented in [27] Lemma 3.15 applies only to the error $E_H^{\mathrm{MS}} = u - I_H^{\mathrm{MS}} u$ because the proof utilises the fact that $E(x_i^\tau) = 0$ for any node of τ. The proof can instead be applied to more general $v \in C(\partial\tau)$ by a simplification of the proof.

Theorem 3.19. *Under Assumption 3.18 let $v \in C(\partial\tau)$, then there exists an extension $\tilde{v} \in H^1(\tau)$ with $\tilde{v} = v$ on $\partial\tau$ that satisfies*

$$|\tilde{v}|^2_{H^1(\tau),\alpha} \lesssim H_\tau^2 \left(\hat{\alpha} \max_{i=1,2,3} \|D_{e_i} v\|^2_{L_\infty(e_i \cap \tau^-)} + \max_{i=1,2,3} \|D_{e_i} v\|^2_{L_\infty(e_i \cap \tau^+)} \right) \ . \tag{3.46}$$

Proof. We assume the geometric situation as in Figure 3-1 where the region that α is high contains x_3^τ. The other cases are analogous. Under Assumption 3.18, we can parametrise τ^- by introducing the local coordinates (t, θ) such that

$$x = tr(\theta)\cos\theta \qquad \text{and} \qquad y = tr(\theta)\sin\theta \tag{3.47}$$

for $t \in [0,1]$ and $\theta \in [0, \beta]$. Then define $\tilde{v}$ explicitly on τ^- by:

$$\tilde{v}(t,\theta) = \left(\frac{\theta}{\beta}\right) v(x_3^\tau + t r_2^- \mathbf{e}_2) + \left(1 - \frac{\theta}{\beta}\right) v(x_3^\tau + t r_1^- \mathbf{e}_1) \ . \tag{3.48}$$

Clearly $\tilde{v}$ coincides with v on $e_i \cap \tau^-$ for each $i = 1, 2$ and

$$\begin{aligned} \frac{\partial \tilde{v}}{\partial x}(t,\theta) = \left(\left(\frac{\theta}{\beta}\right) r_2^- (D_{e_2} v)(x_3^\tau + t r_2^- \mathbf{e}_2) + \left(1 - \frac{\theta}{\beta}\right) r_1^- (D_{e_1} v)(x_3^\tau + t r_1^- \mathbf{e}_1) \right) \frac{\partial t}{\partial x} \\ + \frac{1}{\beta} \left(v(x_3^\tau + t r_2^- \mathbf{e}_2) - v(x_3^\tau) + v(x_3^\tau) - v(x_3^\tau + t r_1^- \mathbf{e}_1) \right) \frac{\partial \theta}{\partial x} \end{aligned} \tag{3.49}$$

with an analogous formula for $\partial\tilde{E}/\partial y$. By exactly the same argument as in [Lemma

3.15 [27]] the first term on the right hand side of (3.49) may be estimated by

$$\frac{H_\tau}{r(\theta)}\left|\cos\theta+\frac{r'(\theta)}{r(\theta)}\sin\theta\right|\max_{i=1,2}\|D_{e_i}v\|_{L_\infty(e_i\cap\tau^-)}\lesssim\frac{H_\tau}{r(\theta)}\max_{i=1,2}\|D_{e_i}v\|_{L_\infty(e_i\cap\tau^-)}\ . \quad (3.50)$$

The second term is then bounded by

$$\frac{|\sin\theta|}{tr(\theta)}tH_\tau\|D_{e_i}v\|_{L_\infty(e_i\cap\tau^-)}\leq\frac{H_\tau}{r(\theta)}\|D_{e_i}v\|_{L_\infty(e_i\cap\tau^-)}$$

since $\left|v(x_3^\tau+tr_2^-\mathbf{e}_i)-v(x_3^\tau)\right|\lesssim tH_\tau\|D_{e_i}v\|_{L_\infty(e_i\cap\tau^-)}$ for $i=1,2$. An analogous procedure can be applied to $\partial\tilde{v}/\partial y$ giving the overall estimate

$$|\nabla\tilde{v}(t,\theta)|\lesssim\frac{H_\tau}{r(\theta)}\max_{i=1,2}\|D_{e_i}v\|_{L_\infty(e_i\cap\tau^-)}\ .$$

Therefore, noting that $dx\ dy=tr^2(\theta)\ d\theta\ dt$, we obtain the estimate on τ^-:

$$\begin{aligned}|\tilde{v}|^2_{H^1(\tau^-),\alpha}=\int_{\tau^-}\hat{\alpha}\,|\nabla\tilde{v}(x,y)|^2\ dx\ dy&=\hat{\alpha}\int_0^1\int_0^\beta|\nabla\tilde{v}(t,\theta)|^2\,tr^2(\theta)\ d\theta\ dt\\&\lesssim H_\tau^2\hat{\alpha}\max_{i=1,2}\|D_{e_i}v\|^2_{L_\infty(e_i\cap\tau^-)}\ .\end{aligned} \quad (3.51)$$

Note that the explicit expansion $\tilde{v}$ on τ^- is constructed to have very precise behaviour on τ^-. For the extension into τ^+ it is sufficient to apply the inverse trace theorem which obtains an extension implicitly. Since τ^+ is a Lipschitz domain, the inverse trace theorem gives an extension $\tilde{v}$ that satisfies (since $\alpha\lesssim 1$ on τ^+)

$$\begin{aligned}|\tilde{v}|^2_{H^1(\tau^+),\alpha}&\lesssim|\tilde{v}|^2_{H^1(\tau^+)}\\&=|\tilde{v}-v(y_1)|^2_{H^1(\tau^+)}\\&\lesssim H_\tau^{-1}\|\tilde{v}-v(y_1)\|^2_{L_2(\partial\tau^+)}+H_\tau\,|\tilde{v}-v(y_1)|^2_{H^1(\partial\tau^+)}\ .\end{aligned} \quad (3.52)$$

Firstly,

$$\|\tilde{v}-v(y_1)\|^2_{L_2(\partial\cap\tau^+)}=\sum_{i=1}^3\|\tilde{v}-v(y_1)\|^2_{L_2(\mathbf{e}_i\cap\tau^+)}+\|\tilde{v}-v(y_1)\|^2_{L_2(\Gamma\cap\tau)}\ . \quad (3.53)$$

Then for $x\in\mathbf{e}_2\cap\tau^+$,

$$\begin{aligned}|v(x)-v(y_1)|&=|v(x)-v(x_2^\tau)+v(x_2^\tau)-v(x_1^\tau)+v(x_1^\tau)-v(y_1)|\\&\lesssim H_\tau\left(\|D_{e_2}v\|_{L_\infty(\mathbf{e}_2\cap\tau^+)}+\|D_{e_3}v\|_{L_\infty(\mathbf{e}_3\cap\tau^+)}+\|D_{e_1}v\|_{L_\infty(\mathbf{e}_1\cap\tau^+)}\right)\ .\end{aligned} \quad (3.54)$$

Therefore,

$$\|\tilde{v} - v(y_1)\|^2_{L_2(\mathbf{e}_2 \cap \tau^+)} \lesssim H_\tau^3 \max_{i=1,2,3} \|D_{e_i} v\|^2_{L_\infty(e_i \cap \tau^+)} \tag{3.55}$$

and the results for $\mathbf{e}_3$ and $\mathbf{e}_1$ are analogous. Consider also that for $x \in \Gamma \cap \tau$,

$$\begin{aligned} |\tilde{v}(x) - v(y_1)| = |\tilde{v}(1,\theta) - \tilde{v}(1,0)| = \left| \frac{\theta}{\beta} \left(v(y_2) - v(y_1) \right) \right| &\leq |(v(y_2) - v(y_1))| \\ &\lesssim H_\tau \max_{i=1,2,3} \|D_{e_i} v\|_{L_\infty(\mathbf{e}_i \cap \tau^+)} \end{aligned}$$

by (3.54). Therefore,

$$\|\tilde{v} - v(y_1)\|^2_{L_2(\Gamma \cap \tau)} \lesssim H_\tau^3 \max_{i=1,2,3} \|D_{e_i} v\|^2_{L_\infty(\mathbf{e}_i \cap \tau^+)} \ . \tag{3.56}$$

Substituting (3.55) and (3.56) back into (3.52) gives

$$H_\tau^{-1} \|\tilde{v} - v(y_1)\|^2_{L_2(\partial\tau^+)} \lesssim H_\tau^2 \max_{i=1,2,3} \|D_{e_i} v\|_{L_\infty(\mathbf{e}_i \cap \tau^+)} \ .$$

Now we consider the second term on the right hand side of (3.52),

$$\begin{aligned} |\tilde{v} - v(y_1)|^2_{H^1(\partial\tau^+)} = |\tilde{v}|^2_{H^1(\partial\tau^+)} &= \sum_{i=1}^{3} |v|^2_{H^1(\mathbf{e}_i \cap \tau^+)} + |\tilde{v}|^2_{H^1(\Gamma \cap \tau)} \\ &\lesssim H_\tau \max_{i=1,2,3} \|D_{e_i} v\|_{L_\infty(\mathbf{e}_i \cap \tau^+)} + |\tilde{v}|^2_{H^1(\Gamma \cap \tau)} \ . \end{aligned}$$

The last term is then bounded by considering that $\theta = \theta(s)$ where s denotes arclength along $\Gamma \cap \tau$. Then

$$\left| \frac{d}{ds} \left\{ \tilde{v}(1, \theta(s)) \right\} \right| = \frac{1}{\beta} \left| v(y_2) - v(y_1) \right| \left| \frac{d\theta}{ds} \right| \lesssim H_\tau \max_{i=1,2,3} \|D_{e_i} v\|_{L_\infty(\mathbf{e}_i \cap \tau^+)} \left| \frac{d\theta}{ds} \right|$$

again using (3.54). Therefore

$$|\tilde{v}|^2_{H^1(\Gamma \cap \tau^+)} \lesssim H_\tau^2 \max_{i=1,2,3} \|D_{e_i} v\|^2_{L_\infty(\mathbf{e}_i \cap \tau^+)} \int_0^{|\Gamma \cap \tau|} \left| \frac{d\theta}{ds} \right|^2 \, ds \lesssim H_\tau \max_{i=1,2,3} \|D_{e_i} v\|^2_{L_\infty(\mathbf{e}_i \cap \tau^+)} \ .$$

Inserting this last estimate into (3.52) and combining with (3.51) gives the required result. □

Now we can apply this theorem to the result in Theorem 3.3 and envisage using the following theorem to prove error estimates for other multiscale methods.

Theorem 3.20. *Under Assumption 3.18 let $v_H \in V_H^{MS}$, then $E_H = u - v_H$ is bounded by:*

$$|E_H|^2_{H^1(\tau),\alpha} \lesssim H_\tau^2 \left(\hat{\alpha} \max_{i=1,2,3} \|D_{e_i} E_H\|^2_{L_\infty(e_i \cap \tau^-)} + \max_{i=1,2,3} \|D_{e_i} E_H\|^2_{L_\infty(e_i \cap \tau^+)} \right) + H_\tau^2 \|f\|^2_{L_2(\tau)} \ . \tag{3.57}$$

Proof. The result follows diretly from combining Theorem 3.19 (using $v = E_H$) with Theorem 3.3. □

Therefore any algorithm that can prove suitably robust edge derivatives on the boundary of an element can extend the robustness to the interior of the elements as well. This result shows how important it is to construct coefficient robust local boundary conditions. Returning to the case in [27] where we have $E_H^{\text{MS}} = u - I_H^{\text{MS}} u$, the authors have done just this as reiterated here in Corollary 3.17. Consequently we get the resulting theorem.

Theorem 3.21. *Let u be the exact solution of Problem 2.2 and suppose $\tau \in \mathcal{T}_H^C(\Omega)$. Then, under Assumption 3.18,*

$$\begin{aligned} |E_H^{MS}|^2_{H^1(\tau),\alpha} \lesssim\ & H_\tau^2 \hat{\alpha}^2 \left(|u|^2_{H^2(\tau^-)} + H_\tau |u|^2_{H^{5/2}(\tau^-)} \right) \\ & + H_\tau^2 \left(|u|^2_{H^2(\tau^+)} + H_\tau |u|^2_{H^{5/2}(\tau^+)} \right) + H_\tau^2 \|f\|^2_{L_2(\tau)} \ . \end{aligned} \tag{3.58}$$

Proof. The proof follows by applying Corollary 3.17 to Theorem 3.20 with $E_H = E_H^{\text{MS}}$ specifically. □

3.1.6 Conforming modification and a global error bound

So far the multiscale method that we have discussed constructs multiscale basis functions on each element separately. The boundary condition on a common edge of two neighbouring elements may not necessarily match. If Γ passes through a common edge $\mathbf{e}$ of τ and τ' then the boundary conditions for the multiscale basis functions are constructed separately on τ and τ' and do not have to match along $\mathbf{e}$. Therefore any basis constructed from these functions may be non-conforming. It is easy to make them continuous however by local averaging of the boundary conditions along an edge before the subgrid solve.

Consider two triangles $\tau = \Delta x_1 x_2 x_3$ and $\tau' = \Delta x_4 x_2 x_3$ that share an edge $\overline{x_2 x_3}$. Firstly the local boundary conditions are calculated on each element to give $\phi_{p,\tau}$ and $\phi_{p,\tau'}$ where

$\phi_{4,\tau} = \phi_{1,\tau'} = 0$. Then the new boundary conditions along $\overline{x_2 x_3}$ are constructed by

$$\frac{\phi_{p,\tau} + \phi_{p,\tau'}}{2}$$

where $p = 1, 2, 3, 4$. After we have averaged the boundary conditions we then extend the basis function into the interior by solving the subgrid problem (Problem 3.1). Doing this for all edges yields a conforming method. Through the use of the triangle inequality it is easy to show that this new boundary condition yields multiscale basis functions that still satisfy Corollary 3.17 and so Theorem 3.21 is still true.

Note however though that the basis functions have slightly larger support as in the example above when $p = 4$ the boundary condition may not be zero along $\overline{x_2 x_3}$ and thus the basis function may be non-zero on τ.

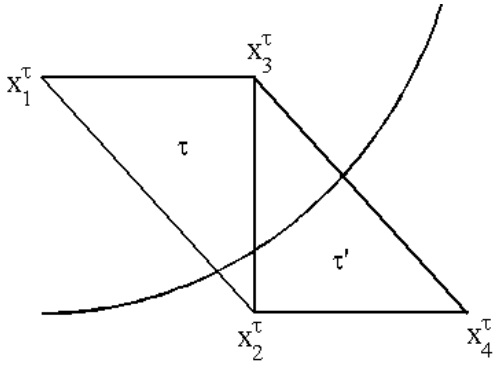


Figure 3-4: *An example of two cut elements that share the edge* $\overline{x_2^\tau x_3^\tau}$.

Combining all of these results for the individual cut elements we obtain a finite element error estimate in the energy norm across the whole domain using Theorem 3.3.

Theorem 3.22. *Let u be the exact solution of Problem 2.2 and suppose τ is a cut element. Then under Assumption 3.18 and assuming $f \in H^{\frac{1}{2}}(\Omega)$ we have for H sufficiently small that*

$$(i) \qquad |u - u_H^{MS}|_{H^1(\Omega),\alpha} \lesssim H \left[H \left| f \right|^2_{H^{\frac{1}{2}}(\Omega)} + \|f\|^2_{L_2(\Omega)} \right]^{\frac{1}{2}}, \tag{3.59}$$

$$(ii) \qquad \|u - u_H^{MS}\|_{H^1(\Omega),\alpha} \lesssim H^2 \left[H \left| f \right|^2_{H^{\frac{1}{2}}(\Omega)} + \|f\|^2_{L_2(\Omega)} \right]^{\frac{1}{2}}. \tag{3.60}$$

Proof. Recall that in Theorem 2.22 the H^{s+2} seminorm of u on Ω_0 was $O(\alpha_0^{-1})$ only on a subset $\widetilde{\Omega_0} \subset \Omega_0$. The boundary of $\widetilde{\Omega_0}$ consists of all the interfaces, Γ, as well as a smooth closed contour $\tilde{\Gamma}$ around all the inclusions Ω_i for $i = 1, ..., m$. For the error

estimate in this theorem H must be small enough so that $\mathcal{T}_H^C(\Omega)$ is contained within $\tilde{\Gamma}$. For Case I (2.8) and using Theorem 3.21 in the finite element error bound in Theorem 3.3 we obtain, after summing over each inclusion,

$$|u - u_H^{\mathrm{MS}}|_{H^1(\Omega),\alpha} \lesssim H^2 \left\{ \hat{\alpha}^2 \sum_{i=1}^{m} \left(|u|^2_{H^2(\Omega_i)} + H\,|u|^2_{H^{5/2}(\Omega_i)} \right) + |u|^2_{H^2(\Omega_0)} + H\,|u|^2_{H^{5/2}(\tilde{\Omega}_0)} + \|f\|^2_{L_2(\Omega)} \right\}. \tag{3.61}$$

Utilising the regularity result in Theorem 2.22 we obtain (3.59). A non-trivial duality argument similar to the one in Theorem 2.60 for the standard finite element method gives the L_2 error estimate (3.60). The proof for Case II (2.9) is similar. □

3.2 Extending to a relative error estimate

The bounds produced in [27] and in Chapter 2 gives estimates that appear to depend on the minimum value of the contrast $\mathcal{A}(x)$ (see Problem 2.2 and Remark 2.8). It would appear to suggest a poor estimate as the minimum of $\mathcal{A}(x)$ tends to zero, however it is in fact the case that the solution blows up in this situation. This means that a relative error bound should be truely independent of the contrast even if the minimum of $\mathcal{A}(x)$ approaches zero. We will prove this result below. For the proof we used an extension of the regularity theory in the appendix of [27].

3.2.1 A regularity result for multiple inclusions

The first stage in proving a relative error estimate is to prove an extension to the regularity theory in [27]. Theorem B.1 in [27] (restated as Theorem 2.22 in this thesis) is given for multiple inclusions in [27]. However the proof is only given for a single inclusion. In this section we restate the theorem but provide the proof for the multiple inclusion case.

Theorem 3.23. *Let Ω be either a smooth C^∞ bounded domain in $\mathbb{R}^2$ or a bounded convex polygon, let Ω contain inclusions Ω_i, i=1,2,...,m, each having a C^∞ boundary, and define $\Omega_0 = \Omega \setminus \cup_{i=1}^m \overline{\Omega}_m$ as described in Definition 2.4. Consider Problem 2.2 and assume that either Case I (2.8) or Case II (2.9) holds. In addition, let $\Gamma = \bigcup_{i=1}^m \Gamma_{0,i}$ and $\tilde{\Gamma}$ denote any closed C^∞ contour in Ω_0, which encloses all the Ω_i and let $\tilde{\Omega}_0$ be the*

domain with boundary $\Gamma \cup \tilde{\Gamma}$. *Then we have*

$$|u|_{H^{s+2}(\Omega_i)} \lesssim \frac{1}{\alpha_i} \|f\|_{H^s(\Omega)}, \text{ for all } s \geq 0, \quad i = 1, 2, ..., m.$$

Moreover

$$|u|_{H^2(\Omega_0)} \lesssim \frac{1}{\alpha_0} \|f\|_{L_2(\Omega)},$$

and

$$|u|_{H^{2+s}(\tilde{\Omega}_0)} \lesssim \frac{1}{\alpha_0} \|f\|_{H^s(\Omega)}, \text{ for all } s \geq 0.$$

The hidden constants depend on the distance of Γ *from* $\partial\Omega$.

Proof. This proof follows a similar style to that found in [27] but with a few key differences that allow it to work for multiple inclusions. We consider only the more complicated case when Ω is a convex polygon (when $\partial\Omega$ is smooth simply take $\widetilde{\Omega}_0 = \Omega_0$). We do not assume one inclusion and instead consider all m inclusions within Ω such as in Figure 2-1. We first consider Case I (2.8) where $\hat{\alpha}$ becomes very large in the inclusions and we explore Case II (2.9) at the end of the proof.

For this proof we denote each interface by $\Gamma_i := \partial\Omega_i$ for $i = 1, ..., m$ and union of the interfaces by $\Gamma := \bigcup_{i=1}^{m} \Gamma_i$. Then we recall two classical regularity results for elliptic boundary value problems. Let $s \geq 0$ and let $\phi \in H^{s+3/2}(\Gamma)$. Then

$$\left\{ \begin{array}{c} \Delta z = w \text{ on } \Omega_i \\ z = \phi \text{ on } \Gamma_i \\ w \in H^s(\Omega_i) \end{array} \right\} \Rightarrow \|z\|_{H^{s+2}(\Omega_i)} \lesssim \|w\|_{H^s(\Omega_i)} + \|\phi\|_{H^{s+\frac{3}{2}}(\Gamma_i)} \tag{3.62}$$

and

$$\left\{ \begin{array}{c} \Delta z = w \text{ on } \Omega_0 \\ z = \phi \text{ on } \Gamma \\ z = 0 \text{ on } \partial\Omega \\ w \in H^s(\Omega_0) \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \|z\|_{H^2(\Omega_0)} \lesssim \|w\|_{L_2(\Omega_0)} + \|\phi\|_{H^{\frac{3}{2}}(\Gamma)} \\ \|z\|_{H^{s+2}(\tilde{\Omega}_0)} \lesssim \|w\|_{H^s(\Omega_0)} + \|\phi\|_{H^{s+\frac{3}{2}}(\Gamma)} \end{array} \right\}. \tag{3.63}$$

Chu, Graham and Hou give suitable references for each of these results in their proof in the appendix of [27]. Now the first step of the proof is to introduce the decomposition

$$u = \hat{u} + \tilde{u} \tag{3.64}$$

where $\hat{u}$ solves the independent Dirichlet problems with homogeneous boundary data

on each Ω_i:

$$\begin{cases} -\alpha_i \Delta \hat{u} &= f \quad \text{on } \Omega_i \\ \hat{u} &= 0 \quad \text{on } \partial\Omega_i \end{cases}$$

for $i = 0, ..., m$. Then from (3.62) and (3.63) we obtain for all $s \geq 0$,

$$\|\hat{u}\|_{H^{2+s}(\Omega_i)} \lesssim \frac{\|f\|_{H^s(\Omega_i)}}{\alpha_i} , \quad \|\hat{u}\|_{H^2(\Omega_0)} \lesssim \frac{\|f\|_{L_2(\Omega_0)}}{\alpha_0} \quad \text{and} \quad \|\hat{u}\|_{H^{2+s}(\tilde{\Omega}_0)} \lesssim \frac{\|f\|_{H^s(\Omega_0)}}{\alpha_0} . \tag{3.65}$$

Thus $\hat{u}$ satisfies the bounds in the statement of the theorem. The remainder of the proof is concerned with obtaining the same bounds for $\tilde{u}$. Since $\tilde{u} = u - \hat{u} \in H_0^1(\Omega)$, it follows tha

$$\begin{cases} \Delta \tilde{u} &= 0 \quad \text{on } \Omega_i \\ \tilde{u} &=: \tilde{v}_i \quad \text{on } \partial\Omega_i \\ \tilde{u} &= 0 \quad \text{on } \partial\Omega \end{cases} \tag{3.66}$$

for $i = 0, ..., m$. As $H_0^1(\Omega)$ is embedded in $C(\Omega)$ then u is continuous and $\hat{u} = 0$ on each interface, consequently $\tilde{u} = u - \hat{u}$ is continuous across each Γ_i. Thus we can define $\tilde{v}_i := \tilde{u}|_{\Gamma_i}$ for $i = 1, ..., m$.

For any suitably smooth v defined on Ω, we let $\partial v_i / \partial n$ denote the normal derivative of v evaluated on Γ with the value taken from within Ω_i, $i = 0, ..., m$ where the normal direction is fixed as outward from Ω_i $i = 1, ..., m$. Then the usual jump relation (3.25) for the solution u of the interface problem, Problem 2.2, reads

$$\frac{\partial u_0}{\partial n} - \alpha_i \frac{\partial u_i}{\partial n} = 0$$

which immediately implies that the function $\tilde{u}$ satisfies the following equation on Γ_i:

$$\frac{\partial \tilde{u}_0}{\partial n} - \alpha_i \frac{\partial \tilde{u}_i}{\partial n} = G_i := \alpha_i \frac{\partial \hat{u}_i}{\partial n} - \frac{\partial \hat{u}_0}{\partial n} . \tag{3.67}$$

This may be readily written:

$$(\mathcal{N}_{0,i} - \alpha_i \mathcal{N}_i)\tilde{v}_i = G_i \tag{3.68}$$

where $\mathcal{N}_i$ denotes the appropriate Dirichlet to Neumann maps on Ω_i. That is, for

$$\begin{cases} \Delta \tilde{v} &= 0 \quad \text{on } \Omega_i \\ \tilde{v} &= \tilde{v}_i \quad \text{on } \partial\Omega_i \end{cases} , \tag{3.69}$$

for $i = 1, ..., m$, $\mathcal{N}_i : H^{s+3/2}(\Gamma_i) \to H^{s+1/2}(\Gamma_i)$ defines the map

$$\mathcal{N}_i(\tilde{v}_i) = \frac{\partial \tilde{v}_i}{\partial n} \ ,$$

recalling that $\tilde{v}_i$ means the limit taken from within Ω_i. The map $\mathcal{N}_{0,i}$ is found for each interface by adding zero boundary data to $\partial\Omega$ and $\Gamma\backslash\Gamma_i$. Since for $i = 1, ..., m$ $\mathcal{N}_i$ has a non-trivial kernel (namely the constant functions on Γ_i denoted $\langle 1 \rangle$), we must study the operator $\mathcal{N}_i$ in the orthogonal complement of this kernel. Thus we introduce the orthogonal projection from $L_2(\Gamma_i)$ onto $\langle 1 \rangle$

$$\mathcal{P}_i v = \frac{1}{|\Gamma_i|} \int_{\Gamma_i} v(s) \ ds$$

for each $i = 1, ..., m$ and $(I - \mathcal{P}_i)$, the orthogonal projection onto

$$L_2(\Gamma_i)^\perp := \{ v \in L_2(\Gamma_i) \mid \mathcal{P}_i v = 0 \} .$$

Then writing

$$\tilde{v}_i = \mathcal{P}_i \tilde{v}_i + (I - \mathcal{P}_i)\tilde{v}_i =: \tilde{c}_i + \tilde{w}_i \ \ ,$$

we can express the jump relations (3.68) as a system in $\langle 1 \rangle \times L_2(\Gamma_i)^\perp$

$$\begin{bmatrix} \mathcal{P}_i(\mathcal{N}_{0,i} - \alpha_i \mathcal{N}_i)\mathcal{P}_i & \mathcal{P}_i(\mathcal{N}_{0,i} - \alpha_i \mathcal{N}_i)(I - \mathcal{P}_i) \\ (I - \mathcal{P}_i)(\mathcal{N}_{0,i} - \alpha_i \mathcal{N}_i)\mathcal{P}_i & (I - \mathcal{P}_i)(\mathcal{N}_{0,i} - \alpha_i \mathcal{N}_i)(I - \mathcal{P}_i) \end{bmatrix} \begin{bmatrix} \tilde{c}_i \\ \tilde{w}_i \end{bmatrix} = \begin{bmatrix} \mathcal{P}_i G_i \\ (I - \mathcal{P}_i) G_i \end{bmatrix} . \tag{3.70}$$

Moreover since $\mathcal{P}_i \mathcal{N}_i = \mathcal{N}_i \mathcal{P}_i$ are null operators on $L_2(\Gamma_i)$, (3.70) can be re-written as

$$\left(\mathbf{P}_i - \alpha_i^{-1} \mathbf{Q}_i \right) \begin{bmatrix} \tilde{c}_i \\ \alpha_i \tilde{w}_i \end{bmatrix} = \begin{bmatrix} \mathcal{P}_i G_i \\ (I - \mathcal{P}_i) G_i \end{bmatrix} \ , \tag{3.71}$$

where

$$\mathbf{P}_i = \begin{bmatrix} \mathcal{P}_i \mathcal{N}_{0,i} \mathcal{P}_i & 0 \\ (I - \mathcal{P}_i)\mathcal{N}_{0,i}\mathcal{P}_i & -\mathcal{N}_i \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_i = \begin{bmatrix} 0 & \mathcal{P}_i \mathcal{N}_{0,i}(I - \mathcal{P}_i) \\ 0 & (I - \mathcal{P}_i)\mathcal{N}_{0,i}(I - \mathcal{P}_i) \end{bmatrix} .$$

We next show that each $\mathbf{P}_i$ is invertible on $\langle 1 \rangle \times L_2(\Gamma_i)^\perp$. Note first that $\mathcal{N}_i$ is invertible on $L_2(\Gamma_i)^\perp$, since the solution to (3.69) is unique up to a constant which has been removed from $L_2(\Gamma_i)^\perp$. To analyse $\mathcal{P}_i \mathcal{N}_{0,i} \mathcal{P}_i$ consider the boundary value problem:

$$\Delta \eta_i = 0 \quad \text{in } \Omega_0, \quad \text{with } \ \eta_i = 1 \ \text{ on } \Gamma_i \quad \text{and} \quad \eta_i = 0 \quad \text{on } \partial\Omega \cup (\Gamma \backslash \Gamma_i) , \tag{3.72}$$

which has a unique solution $\eta_i \in H^2(\Omega_0)$. The linear operator $\mathcal{P}_i\mathcal{N}_{0,i}\mathcal{P}_i$ operates on $\langle 1\rangle$ as multiplication by the scalar

$$\gamma_i := \mathcal{P}_i\left[\frac{\partial\eta_i}{\partial n}\right] = \frac{1}{|\Gamma_i|}\int_{\Gamma_i}\frac{\partial\eta_i}{\partial n}\,ds\ .$$

To see this consider that for any $c \in \langle 1\rangle$,

$$\mathcal{P}_i\mathcal{N}_{0,i}\mathcal{P}_i c = \mathcal{P}_i\mathcal{N}_{0,i}c = \mathcal{P}_i\left[\frac{c\partial\eta_i}{\partial n}\right] = c\mathcal{P}_i\left[\frac{\partial\eta_i}{\partial n}\right] = c\gamma_i\ .$$

Note that the scalar γ_i does not vanish, since (by (3.72)),

$$\gamma_i\,|\Gamma_i| = \int_{\Gamma_i}\frac{\partial\eta_i}{\partial n}\,ds = \int_{\partial\Omega_0}\eta_i\frac{\partial\eta_i}{\partial n}\,ds = \int_{\Omega_0}\nabla\cdot(\eta_i\nabla\eta_i)\,dx = \int_{\Omega_0}|\nabla\eta_i|^2\,dx > 0\ .$$

Moreover the linear operator $(I-\mathcal{P}_i)\mathcal{N}_{0,i}\mathcal{P}_i$ operates on $\langle 1\rangle$ as multiplication by the function $\rho_i := (I-\mathcal{P}_i)(\partial\eta_i/\partial n) = \partial\eta_i/\partial n - \gamma_i \in L_2(\Gamma_i)^\perp$. Again for any $c \in \langle 1\rangle$

$$(I-\mathcal{P}_i)\mathcal{N}_{0,i}\mathcal{P}_i c = \mathcal{N}_{0,i}\mathcal{P}_i c - \mathcal{P}_i\mathcal{N}_{0,i}\mathcal{P}_i c = \mathcal{N}_{0,i}c - c\gamma_i = c\,(\partial\eta_i/\partial n - \gamma_i) = c\rho_i\ .$$

Hence

$$\mathbf{P}_i = \begin{bmatrix}\gamma_i & 0\\ \rho_i & -\mathcal{N}_i\end{bmatrix} \quad \text{and} \quad \mathbf{P}_i^{-1} = \begin{bmatrix}\gamma_i^{-1} & 0\\ \gamma_i^{-1}\mathcal{N}_i^{-1}\rho_i & -\mathcal{N}_i^{-1}\end{bmatrix}\ .$$

Now combining (3.62) and (3.63) with the Trace Theorem we obtain that $\mathcal{N}_i : L_2(\Gamma_i)^\perp \cap H^{s+3/2}(\Gamma_i) \to L_2(\Gamma_i)^\perp \cap H^{s+1/2}(\Gamma_i)$ is a bounded operator and has a bounded inverse (since $\mathcal{N}_i$ is a bijective bounded linear operator). We also have that $\mathcal{N}_{0,i} : H^{s+3/2}(\Gamma) \to H^{s+1/2}(\Gamma)$ is bounded and $\mathbf{P}_i^{-1}\mathbf{Q}_i$ is a bounded operator on $\langle 1\rangle \times H^{s+3/2}(\Gamma_i)$ for each $i = 1, ..., m$. Then we have that

$$\left\|\mathbf{P}_i^{-1}\begin{bmatrix}\mathcal{P}_i G_i\\ (I-\mathcal{P}_i)G_i\end{bmatrix}\right\|_{\langle 1\rangle\times H^{s+3/2}(\Gamma_i)} \lesssim \|G_i\|_{H^{s+1/2}(\Gamma_i)}\ . \tag{3.73}$$

Hence, considering (3.71) and that we are examining Case I (2.8), we have for sufficiently large $\hat{\alpha}$ that

$$\begin{aligned}\max\left\{|\tilde{c}_i|\ ,\ \alpha_i\|\tilde{w}_i\|_{H^{s+3/2}(\Gamma_i)}\right\} &\lesssim \|G_i\|_{H^{s+1/2}(\Gamma_i)}\\ &\le \alpha_i\left\|\frac{\partial\hat{u}_i}{\partial n}\right\|_{H^{s+1/2}(\Gamma_i)} + \left\|\frac{\partial\hat{u}_0}{\partial n}\right\|_{H^{s+1/2}(\Gamma_i)}\\ &\lesssim \alpha_i\|\hat{u}\|_{H^{2+s}(\Omega_i)} + \|\hat{u}\|_{H^{2+s}(\tilde{\Omega}_0)}\\ &\lesssim \|f\|_{H^s(\Omega)}\ ,\end{aligned} \tag{3.74}$$

using the definition of each G_i in (3.67), then the trace theorem and finally (3.65).

Now recall that $\tilde{u}$ is harmonic on each Ω_i and that $\tilde{u}|_{\Gamma_i} =: \tilde{v}_i = \tilde{c}_i + \tilde{w}_i$, where $\tilde{c}_i \in \mathbb{R}$. Thus, define $\tilde{u}_i$ on Ω_i for $i = 1, .., m$ by

$$\begin{cases} \Delta \tilde{u}_i &= 0 \quad \text{on } \Omega_i \\ \tilde{u} &= \tilde{w}_i \quad \text{on } \Gamma_i \end{cases},$$

and by uniqueness we have, $\tilde{u} = \tilde{c}_i + \tilde{u}_i$ on Ω_i, $i = 1, .., m$. Thus using (3.62) and then (3.74), we have for all $s \geq 0$,

$$|\tilde{u}|_{H^{2+s}(\Omega_i)} = |\tilde{u}_i|_{H^{2+s}(\Omega_i)} \lesssim \|\tilde{w}_i\|_{H^{s+3/2}(\Gamma_i)} \lesssim \frac{1}{\alpha_i} \|f\|_{H^s(\Omega)} \ . \tag{3.75}$$

Combining (3.75) with the first inequality in (3.65) yields the first required estimate on each Ω_i. To obtain the estimates on Ω_0 we note that (3.74) implies that each $\tilde{v}_i$ satisfies $\|\tilde{v}_i\|_{H^{s+3/2}(\Gamma_i)} \lesssim \|f\|_{H^s(\Omega)}$ and hence the required results follow from (3.63).

Finally we remark that Case II, where $\hat{\alpha} = \alpha_0 \to \infty$ and $\alpha_i \leq K$, is easier to prove. In this case the analysis of $\hat{u}$ is unchanged but in the analysis of each $\tilde{v}_i$ we obtain

$$(\alpha_0 \mathcal{N}_{0,i} - \mathcal{N}_i)\tilde{v}_i = G_i := \frac{\partial \hat{u}_i}{\partial n} - \alpha_0 \frac{\partial \hat{u}_0}{\partial n}$$

instead of (3.68). Here, $\mathcal{N}_{0,i}$ is understood to be the Dirichlet to Neumann map that results fro

$$\begin{cases} \Delta \tilde{u}_i &= 0 \quad \text{on } \Omega_i \\ \tilde{u} &= \tilde{v}_i \quad \text{on } \Gamma_i \\ \tilde{u} &= 0 \quad \text{on } \partial\Omega \cup (\Gamma \backslash \Gamma_i) \end{cases}.$$

Since $\mathcal{N}_{0,i}$ is invertible the estimate for $\tilde{v}_i$ can then be obtained by premultiplying by $\alpha_0^{-1}\mathcal{N}_{0,i}^{-1}$ and letting α_0 get sufficiently large, thus avoiding the projection procedure. □

3.2.2 A relative error estimate for the high-contrast interface problem

Now we can consider the unscaled problem as mentioned in Remark 2.8. Here the permeability field $\mathcal{A}(x)$ is allowed to tend to zero on some inclusions. In order to prove robustness with respect to the contrast parameter in this case we first need a lemma providing a lower bound for the H^2-seminorm of u in each Ω_i, under certain assumptions.

Lemma 3.24. *Suppose the permeability field $\mathcal{A}(x)$ satisfies the assumptions of Theorem 3.23. Suppose also that $supp(f) \cap \Omega_{\min} \neq \emptyset$ where $supp(\cdot)$ is the support of a function and $\Omega_{\min}$ is the inclusion with minimum coefficient $\mathcal{A}_{\min}$. Then we have*

$$\frac{\|f\|_{L_2(\Omega_{\min})}}{\min_i \mathcal{A}_i} \lesssim \max_i |u|_{H^2(\Omega_i)} . \tag{3.76}$$

The hidden constants depend on the distance of the interface Γ from the boundary $\partial\Omega$.

Proof. First we label the inclusion with minimum coefficient $\mathcal{A}_{\min}$ as $\Omega_{\min}$. Then consider the scaled problem

$$\int_\Omega \alpha \nabla u \cdot \nabla v \ \ dx = \int_\Omega \frac{fv}{\mathcal{A}_{\min}} \ \ dx$$

where $\alpha = \mathcal{A}/\mathcal{A}_{\min}$. Now we consider the decomposition of u as in the proof of Theorem 3.23. Let

$$u = \hat{u} + \tilde{u}$$

where $\hat{u}$ solves the independent Dirichlet problems

$$\begin{cases} -\alpha_i \Delta \hat{u} &= \frac{f}{\mathcal{A}_{\min}} \quad \text{on } \Omega_i \\ \hat{u} &= 0 \quad \text{on } \partial\Omega_i \end{cases} \tag{3.77}$$

for each $i = 0, ..., m$. Then, (3.77) defines a bijective solution operator $\mathcal{G} : L_2(\Omega_{\min}) \to H^2(\Omega_{\min})$ such that $\mathcal{G}(f) = u$ and,

$$|\hat{u}|_{H^2(\Omega_{\min})} = \frac{1}{\mathcal{A}_{\min}} |\mathcal{G}(f)|_{H^2(\Omega_{\min})} \ . \tag{3.78}$$

Trivially this is bounded below by the estimate

$$|\hat{u}|_{H^2(\Omega_{\min})} \geq \|\Delta u\|_{L_2(\Omega_{\min})} = \frac{1}{\mathcal{A}_{\min}} \|f\|_{L_2(\Omega_{\min})} \ .$$

From (3.75) and analogously for Ω_0 (using the first estimate in (3.63)) we have the bounds

$$|\tilde{u}|_{H^2(\Omega_i)} \lesssim \frac{1}{\alpha_i \mathcal{A}_{\min}} \|f\|_{L_2(\Omega)} \tag{3.79}$$

for $i = 0, ..., m$. Now we utilise results (3.78) and (3.79) in the inverse triangle inequality to get

$$\begin{aligned}
|u|_{H^2(\Omega_{\min})} &\geq |\hat{u}|_{H^2(\Omega_{\min})} - |\tilde{u}|_{H^2(\Omega_{\min})} \\
&\gtrsim \frac{1}{\mathcal{A}_{\min}} \|f\|_{L_2(\Omega_{\min})} - \frac{1}{\hat{\alpha}} \frac{1}{\mathcal{A}_{\min}} \|f\|_{L_2(\Omega)} \\
&= \frac{1}{\mathcal{A}_{\min}} \left\{ \|f\|_{L_2(\Omega_{\min})} - \frac{1}{\hat{\alpha}} \|f\|_{L_2(\Omega)} \right\} \\
&\gtrsim \frac{1}{\mathcal{A}_{\min}} \|f\|_{L_2(\Omega_{\min})}
\end{aligned}$$

when $\hat{\alpha}$ is sufficiently large to ensure that $\|f\|_{L_2(\Omega)} / \hat{\alpha} \leq \frac{1}{2} \|f\|_{L_2(\Omega_{\min})}$. Thus,

$$\frac{\|f\|_{L_2(\Omega_{\min})}}{\min_i \mathcal{A}_i} \lesssim |u|_{H^2(\Omega_{\min})} \leq \max_i |u|_{H^2(\Omega_i)} \ .$$

□

Remark 3.25. *Note that Lemma 3.24 shows that (2.14) in Remark 2.8 holds since for* $\mathcal{A}_{\min} < 1$

$$\mathcal{A}_{min}^{-\frac{1}{2}} \leq \mathcal{A}_{min}^{-1} \leq C \max_i |u|_{H^2(\Omega_i)}$$

and thus we obtain the corresponding robust finite element error estimates. We clarify these robust estimates in the following theorem for both the case of the standard finite element method as well as the multiscale finite element method in [27].

Theorem 3.26. *Firstly suppose that* u *solves*

$$\begin{cases} \int_\Omega \nabla u \cdot \mathcal{A} \nabla v \ \ dx = \int_\Omega f v \ \ dx & \text{for any } v \in H_0^1(\Omega) \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{3.80}$$

and that the rescaled permeability field $\alpha = \mathcal{A}/\mathcal{A}_{\min}$ *satisfies the Assumptions 2.20, 2.15 and 2.25. Suppose also that* $\mathcal{A}_{\min} := \min_i \mathcal{A}_i < 1$ *Then the standard finite element error* $u - u_H$ *is bounded by*

$$\textit{(i)} \ \frac{|u - u_H|_{H^1(\Omega),\mathcal{A}}}{\max_i |u|_{H^2(\Omega_i)}} \lesssim \frac{(1 + \eta_H)\, H^{\frac{1}{2}-\epsilon} \|f\|_{H^s(\Omega)}}{\|f\|_{L_2(\Omega_{\min})}} \ , \tag{3.81}$$

$$\textit{(ii)} \ \frac{|u - u_H|_{L_2(\Omega),\mathcal{A}}}{\max_i |u|_{H^2(\Omega_i)}} \lesssim \frac{(1 + \eta_H)^2\, H^{1-2\epsilon} \|f\|_{H^s(\Omega)}}{\|f\|_{L_2(\Omega_{\min})}} \tag{3.82}$$

for $s > 0$. *Suppose instead that the rescaled permeability field* α, f *and* H *satisfy the assumptions of Theorem 3.22. Then the multiscale finite element error* $u - u_H^{MS}$ *from [27] is bounded by*

$$(i)\ \frac{|u-u_H^{MS}|_{H^1(\Omega),\mathcal{A}}}{\max_i |u|_{H^2(\Omega_i)}} \lesssim \frac{H\left[H\,|f|^2_{H^{\frac{1}{2}}(\Omega)} + \|f\|^2_{L_2(\Omega)}\right]^{\frac{1}{2}}}{\|f\|_{L_2(\Omega_{\min})}}\ , \tag{3.83}$$

$$(ii)\ \frac{|u-u_H^{MS}|_{L_2(\Omega),\mathcal{A}}}{\max_i |u|_{H^2(\Omega_i)}} \lesssim \frac{H^2\left[H\,|f|^2_{H^{\frac{1}{2}}(\Omega)} + \|f\|^2_{L_2(\Omega)}\right]^{\frac{1}{2}}}{\|f\|_{L_2(\Omega_{\min})}}\ . \tag{3.84}$$

where $\Omega_{\min}$ *is the inclusion with coefficient* $\mathcal{A}_{\min} = \min_i \mathcal{A}_i$.

Proof. Using the definition of the energy norm we have

$$|v|_{H^1(\Omega),\mathcal{A}} \leq \mathcal{A}_{\min}^{\frac{1}{2}}\, |v|_{H^1(\Omega),\alpha}$$

for any function $v \in H^1(\Omega)$. Then applying Lemma 3.24 we obtain

$$\frac{|v|_{H^1(\Omega),\mathcal{A}}}{\max_i |u|_{H^2(\Omega_i)}} \leq \mathcal{A}_{\min}^{3/2} \frac{|v|_{H^1(\Omega),\alpha}}{\|f\|_{L_2(\Omega_{\min})}}\ . \tag{3.85}$$

Then substituting in the result of Theorem 2.58 and Theorem 2.60 respectively into (3.85) to obtain the standard finite element bounds (3.81) and (3.82). Note that f is replaced by $f/\mathcal{A}_{\min}$ in Theorem 2.58 and Theorem 2.60 for the unscaled problem and we note that $\mathcal{A}_{\min}^{1/2} \leq 1$.

The results (3.83) and (3.84) for the multiscale finite element method follow from Theorem 3.22 using a similar substitution into (3.85).

□

3.3 Summary

In this chapter we have given a review of the multiscale finite element method by Chu, Graham and Hou in [27] conveying the ideas for greater understanding but also generalising some of the results. In Section 3.1.1 we explored a key idea behind proving contrast independent error estimates for multiscale finite element methods in general, not only the method presented in [27]. We then gave an overview of the method of proof in [27] so as to convey the ideas for proving a contrast independent finite element error and leave the technical details to following sections. We showed that the artificial local boundary condition is simple to calculate by the solution of a small linear system in Section 3.1.2 but then the analysis of the finite element error is highly

complicated. The analysis was done by examining properties of the true solution in Section 3.1.3, considering the error on the boundary of a cut element in Section 3.1.4 and then extending to the interior error in Section 3.1.5. In Section 3.1.6 all the previous results were brought together to demonstrate how the creation of coefficient robust local boundary conditions leads to multiscale basis functions, which in turn produces a contrast robust finite element error that converges at the same rate as a smooth coefficient $\mathcal{A}$ (i.e. $O(H)$ in the energy norm). The new result in these previous sections comes from giving a more accessible view of the work in [27] but also aiming it towards more general multiscale finite element methods that have multiscale basis functions, this was done by considering the key idea behind coefficient robust error estimates but also by generalising the interior error result in Section 3.1.5 to any function on the boundary and not just the error between the solution and the nodal interpolant. Whilst not a complete generalisation it does present some steps towards analysing other multiscale methods, for example the adaptive method presented in Chapter 4.

Much of the new work came in Section 3.2.1 were we extended the proof of the regularity result in [27] to multiple inclusions and using that result, created a relative error estimate in Section 3.2.2. The relative estimate allows us to see how the error estimate depends explicitly on $\mathcal{A}_{\min}$ as $\mathcal{A}_{\min} \to 0$ as the ellipticity is lost.

What is apparent is that these apriori local boundary conditions are difficult to find for general coefficients $\mathcal{A}$. In the next chapter we consider a method to find the local boundary conditions iteratively.

4

The adaptive multiscale finite element method

In Chapter 2 we showed that, when a mesh did not resolve the interfaces, the energy norm error for the second order elliptic interface Problem (2.1) was at best of $O(H^{\frac{1}{2}})$ with elements of size H. In Chapter 3 we showed how this could be improved to $O(H)$ in the energy norm through finding local boundary conditions for a subgrid problem (Problem 3.1) to obtain multiscale basis functions that give a better approximation. While the multiscale method is straightforward to implement, it only applies to interface problems and makes strong assumptions about how the interface cuts through an element (see Assumption 3.5). The goal set out in this chapter is to develop a method that can find the boundary conditions to the local problems automatically and with any geometry and work for general heterogeneous elliptic problems (not just interface problems). In Section 4.1 we demonstrate why these local boundary conditions are key to finding a coefficient independent error estimate and why it is important to find so called 'good' local boundary conditions.

This chapter introduces a multiscale method that removes the need to know these local boundary conditions a priori. Instead this new adaptive multiscale method seeks to iterate several times from initial local boundary conditions and improve them to get multiscale basis functions that approximate the solution well.

Adaptivity normally takes one of several forms; h-adaptivity seeks to locally refine the size and number of elements within a mesh to improve convergence around parts of the domain. Similar to this, r-adaptivity moves the location of mesh nodes and consequently changes the shape of the mesh to better approximate the solution. Another type, p-adaptivity, involves increasing the order of the polynomials used in the test functions so that the test space better approximates the solution space. The problem with these methods is that the h- and r-adaptivities involve a lot of effort re-meshing a domain if coefficients change and in the h- and p- cases the size of the global matrix system can become large as more and more smaller elements are introduced or the order of polynomial increases. The p-adaptivity has the difficulty of knowing what order of

polynomials to use to sufficiently approximate the solution, for example if the solution is continuous but has a sudden jump in gradient then the degree of polynomial would have to be very high to capture the kink in the solution. The adaptive multiscale finite element method is different from these forms of adaptivity but is most like p-adaptivity. The original mesh remains fixed and the shape of the basis functions change, however, the basis functions for the adaptive multiscale method are not necessarily polynomial in shape. Instead they solve a local homogeneous version of the underlying problem where the local boundary conditions adapt to the fine scale features of the solution iteratively. The importance of finding the so called 'good' boundary conditions is discussed in Section 4.1 but the idea is that these 'good' boundary conditions allow recovery of the true solution without pollution by the coefficient $\mathcal{A}(x)$. It is important to note as well that in the adaptive method used here, this process is all automatic with no input from the user or error indicators describing where to adapt.

The adaptive multiscale method used here has its origins in the paper "An adaptive local-global multiscale finite volume element method for two-phase flow simulations" by Durlofsky, Efendiev and Ginting [36] where it was introduced and applied to two phase flow through porous media in 2D with a finite volume method. The method is far more powerful than demonstrated in [36]. This chapter seeks to give a proper description of both the EDG1 adaptive local-global multiscale finite element method (EDG1 ALG-MsFEM) and the EDG2 ALG-MsFEM, termed the "conforming" and "non-conforming" ALG-MsFEM in [36] respectively, setting it to a much more general context. The convergence rate of each will be numerically demonstrated and it will be shown how the EDG2 method is far superior to the EDG1 method.

The chapter will also introduce a modification to the EDG2 ALG-MsFEM which significantly improves its convergence. The chapter will start by a general description of the idea behind basis function iteration. Then we will describe the iterative process for a particular element of a finite element mesh. The chapter will then move on to show some of the properties of the method as well as a general description of the framework to encompass both the EDG1 and EDG2 methods. Finally the chapter will end by examining numerical convergence results for the method and showing how powerful it is not only when the mesh does not align with the interfaces in the domain but also when the interfaces are not smooth and when the interfaces get close to the boundary of the domain. Finally we use the adaptive multiscale finite element method for some model problems related to porous media flow in the case where the permeability field $\mathcal{A}(x)$ is a random field.

4.1 The idea of 'good' local boundary conditions

Traditionally the h-version of the finite element method uses basis functions that are a fixed polynomial order on each element. For example if they are nodal (i.e. the function ϕ_i is 1 at the node n_i and 0 at all other nodes) and linear on each element then we get the usual set of 'hat' functions.

However in this section we show that by solving a local homogeneous problem on each element then we can get a more intelligent set of basis functions that immediately allows error estimates that are optimal and independent of the contrast. When solving the local homogeneous problems it is important to choose boundary conditions that will lead to good approximations of the true solution $u(x)$. A useful exercise is to ask the question: If the true solution u were known, how should the local boundary conditions be chosen? Let us consider this question in 1D.

Example 4.1. *Assume that we have a function $u(x) \in C([0,1])$ such that $u(0) \neq u(1)$. Then we define the basis functions*

$$\phi_0(x) = \frac{u(x) - u(1)}{u(0) - u(1)} \quad \text{and} \quad \phi_1(x) = \frac{u(x) - u(0)}{u(1) - u(0)} \ .$$

Now if we consider the interpolant

$$(Iu)(x) = u(0)\phi_0(x) + u(1)\phi_1(x) \ ,$$

then it has the property that

$$\begin{aligned}(Iu)(x) &= u(0)\left(\frac{u(x) - u(1)}{u(0) - u(1)}\right) + u(1)\left(\frac{u(x) - u(0)}{u(1) - u(0)}\right) \\ &= \frac{u(x)\left(u(0) - u(1)\right) - u(0)u(1) + u(1)u(0)}{u(0) - u(1)} = u(x) \ .\end{aligned}$$

So by using the u-dependent basis functions ϕ_0, ϕ_1 the interpolant Iu can recover the true solution u from just its values at the end points.

Now consider using the analogue of the example above on a 2D triangular element τ of diameter H_τ and with vertices $\{x_i\}_{i=1}^3$. Then we could define u-dependent basis

functions as

$$\phi_j \mid_{[x_i,x_{i+1}]} = \begin{cases} \dfrac{u(x)-u(x_{i+1})}{u(x_i)-u(x_{i+1})} & \text{if } j=i \\ \dfrac{u(x)-u(x_i)}{u(x_{i+1})-u(x_i)} & \text{if } j=i+1 \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, 2, 3$. Then assuming u has differing values at x_1, x_2 and x_3, from the previous example we know that the nodal interpolant $I_H u = \sum_{i=1}^{3} u(x_j)\phi_j$ also recovers u on $\partial\tau$. The basis functions are then extended harmonically into τ by solving

$$A(\phi_j, v) = 0 \qquad \text{for any } \; v \in H_0^1(\tau) \; .$$

This is significant because if we generalise Theorem 3.3 (Theorem 2.2 [27]) for general bounded and coercive bilinear forms then, as the following lemma shows, we obtain an error on each element that is robust with respect to the contrast in the coefficient $\alpha \in L_\infty(\Omega)$ when using the bilinear form

$$A_\Omega = \int_\Omega \alpha \nabla u \cdot \nabla v \; .$$

Conjecture 4.2. *If we consider the bilinear form $A_\Omega(u,v) = \int_\Omega \mathcal{A}\nabla u \cdot \nabla v$ on $H_0^1(\Omega)$ and any L_∞ coefficient $\mathcal{A}$ bounded away from zero then*

$$C_A := (1/\min_{x\in\Omega} \mathcal{A}(x))^{\frac{1}{2}} \; .$$

Consequently if E_H vanishes on $\partial\tau$ and τ had diameter H then,

$$C_A^{-1} \, |E_H|_{E(\tau)} \leq CH \, \|f\|_{L_2(\tau)} \; . \tag{4.1}$$

Importantly this error estimate is completely independent of the contrast in the coefficient $\mathcal{A}(x)$ (C_A is a measure of poor ellipticity of A_Ω).

Lemma 4.3. *Let $A_D(\cdot,\cdot)$ be the local version of the bilinear form $A(\cdot,\cdot)$ on a Lipschitz subdomain D of Ω and denote $|\cdot|_{E(D)} = A_D(\cdot,\cdot)^{\frac{1}{2}}$ as the corresponding energy norm. Assume that there exists a constant C_A such that $|v|_{H^1(D)} \leq C_A \, |v|_{E(D)}$ independent of*

the domain D and argument v. Suppose that $v \in H^1(D)$ satisfies

$$A_D(v, w) = \int_D fw \qquad \text{for any } w \in H_0^1(D) \ .$$

Then for any $\tilde{v} \in H^1(D)$ such that the trace of $\tilde{v} - v$ vanishes on ∂D,

$$|v|_{E(D)} \leq |\tilde{v}|_{E(D)} + CC_A diam(D) \, \|f\|_{L_2(D)} \tag{4.2}$$

where C is independent of v, $\tilde{v}$, the diameter of D and A.

Proof. Let v^* be the unique solution of the problem

$$A_D(v^*, w) = 0 \qquad \text{for any } w \in H_0^1(D) \tag{4.3}$$

such that the trace of $v^* - v$ vanishes on ∂D. Then $v - v^* \in H_0^1(D)$ and

$$A_D(v - v^*, w) = \int_D fw \ \ dx \qquad \text{for any } w \in H_0^1(D) \ .$$

Therefore

$$\begin{aligned} |v - v^*|^2_{E(D)} = A_D(v - v^*, v - v^*) = \int_D f(v - v^*) \ \ dx \\ \leq \|f\|_{L_2(D)} \, \|v - v^*\|_{L_2(D)} \\ \leq CC_A \text{diam}(D) \, \|f\|_{L_2(D)} \, |v - v^*|_{E(D)} \ , \end{aligned}$$

where the last step uses the Poincaré-Friedrichs inequality and the assumption that $|\cdot|_{H^1(D)} \leq C_A \, |\cdot|_{E(D)}$. After dividing both sides by $|v - v^*|_{E(D)}$ and using the inverse triangle inequality we get

$$|v|_{E(D)} \leq |v^*|_{E(D)} + CC_A \text{diam}(D) \, \|f\|_{L_2(D)} \ .$$

However (4.3) implies minimality of the energy norm of v^* so $|v^*|_{E(D)} \leq |\tilde{v}|_{E(D)}$ for all $\tilde{v}$ satisfying the same boundary conditions as v and the result follows. □

Since $u - I_H u = 0$ on $\partial\tau$ we obtain the following corollary.

Corollary 4.4. *Under the same conditions as Lemma 4.3 with domain $D = \tau$, if $E_H = u - I_H u$ vanishes on $\partial\tau$ then,*

$$\frac{1}{C_A} \, |E_H|_{E(\tau)} \leq C diam(\tau) \, \|f\|_{L_2(\tau)} \tag{4.4}$$

where C is independent of v, $\tilde{v}$, the diameter of τ and A.

Proof. The proof follows from Lemma 4.3 using $v = E_H$ and $\tilde{v} = 0$. □

The previous lemma and corollary prove Conjecture 4.2. They show that if we knew the exact solution along the edges of the elements of a mesh $\mathcal{T}_H$ then we could define a suitable set of multiscale basis functions which would give a solution with a coefficient independent error estimate. Therefore we define 'good' local boundary conditions as ones that result in the interpolant being close to the true solution.

4.2 The idea of basis function iteration

In this section we try to convey the fundamental idea that makes the adaptive multiscale method in this chapter work. In the previous section we showed how to construct basis functions that allow coefficient independent error estimates but it made the assumption of knowing the true solution $u(x)$ a priori on the boundaries of the elements. This is not known normally but instead consider starting with any local boundary conditions (e.g. linear conditions) and iterating to get closer to these 'ideal' local boundary conditions.

Supposing we have an initial set of boundary conditions that could be very far from the 'ideal' ones that we want. By solving a local homogeneous problem on a larger oversampled domain around an element we can reduce the effect of poor initial local boundary conditions on the solution in the interior of this domain, and specifically within the original unextended element. We then need to take a linear combination of these local solutions to create basis functions on the element (with the aim that if the interpolant reproduces the true solution (see Section 4.1) we can use Corollary 4.4 to obtain a robust error estimate). The linear combination process will be explained further in Section 4.4. Using these basis functions we can solve the variational multiscale finite element problem (Problem 5.3) on the whole domain to get an approximate solution. We are then able to repeat this whole process by generating new local boundary conditions from the approximate solution with the aim that they will converge to the 'ideal' ones resulting from the true solution.

4.3 The iterative cycle

In this section we will describe the specifics of how these multiscale basis functions are created. The process described forms the main iterative step of the adaptive multiscale

method framework.

4.3.1 Inputs to the iterative step

The iterative step takes as its input the coefficient, an initial approximation to the solution, a particular element τ currently being processed and the corresponding extended element $\tilde{\tau}$. As described in the previous section we are trying to use a larger domain around τ to dilute the effect of an inaccurate local boundary condition, for now we make the assumption that such an extension $\tilde{\tau} \supseteq \tau$ exists.

The exact implications for the choice of extended element needs more investigation. It is worth noting that the extended triangle could be set to τ itself, as is the case for the EDG1 ALG-MsFEM (see Section 4.5.2) but as expected this does not allow a starting local boundary condition to improve. Previous work by Nolen, Papanicolaou and Pironneau [72] suggests that the gap between $\partial\tau$ and $\partial\tilde{\tau}$ should be at least one coarse mesh element or more, particularly in the case of periodic and random coefficient $\mathcal{A}$ which they numerically suggest needs four or more layers. Note that the oversampling may require the whole domain which destroys the possibility of having a local process. Further investigation is needed because the numerical results in Section 4.6 for the Adaptive Multiscale FEM show good convergence even in the case of a random coefficient for just one layer of oversampling. Experimental results suggest that it is only the number of iterations required that is affected by increasing the width of the gap between $\partial\tau$ and $\partial\tilde{\tau}$ and so adapting the size of the gap is unnecessary.

An example of an extended element $\tilde{\tau}$ is shown in Figure 4-1. This example $\tilde{\tau}$ has the property that the data required for the local boundary conditions on $\partial\tilde{\tau}$ can be aquired from the data on the edges of the mesh $\mathcal{T}_H(\Omega)$.

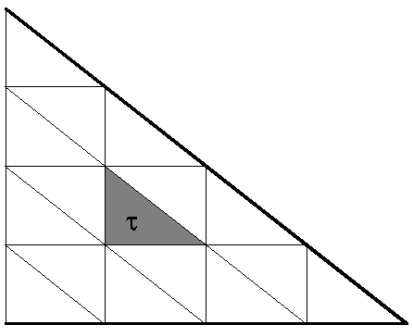


Figure 4-1: *An example of an extended element $\tilde{\tau}$ around an element τ. The figure demonstrates how this can line up with the mesh $\mathcal{T}_H(\Omega)$.*

We now have the components necessary for the iterative step assuming we have been given an approximate solution u. Now we examine the local homogeneous problem that leads to the multiscale basis functions. The next section describes how the boundary data for the local problem is calculated to reflect the features of the true solution.

4.3.2 The adaptive multiscale method edge mapping function

The key feature of the ALG-MsFEM methods in [36] is that the boundary condition for the local problem preserves the fine scale features of the current approximation. Note however that if the approximation is poor then it will remain poor unless oversampling is used as in Figure 4-1 and then the fine scale features still enter in to the approximation but it can converge to the true solution. This means that the basis functions can capture the fine scale features of the solution while solving the finite element problem on a coarse mesh. These conditions are found by using a 1-dimensional map $\mathcal{P}_e$ along the edges of the triangle τ.

Definition 4.5. *Let $e = \{a + t(b-a) \mid t \in [0,1]\}$ be an edge that connects a to b. Then define $\mathcal{P}_e : C(e) \to \mathbb{R}$ for any $u \in C(e)$ and $x \in e$ by*

$$\mathcal{P}_e u(x) = \begin{cases} \frac{u(x)-u(a)}{u(b)-u(a)} & \text{if } u(a) \neq u(b) \\ \Psi_e(x) + \frac{u(x)-u(a)}{2u(a)} & \text{if } u(a) = u(b) \neq 0 \\ \Psi_e(x) & \text{if } u(a) = u(b) = 0 \end{cases},$$

where $\Psi_e(x) = (x-a)/(b-a)$ is the linear function satisfying $\Psi_e(a) = 0$ and $\Psi_e(b) = 1$.

Proposition 4.6. *The function $\mathcal{P}_e$ is actually a projection on $C(e)$.*

Proof. For an edge e and function $u \in L_\infty(e)$, $\mathcal{P}_e u$ maps to the values $\mathcal{P}_e u(a) = 0$ and $\mathcal{P}_e u(b) = 1$ with behaviour on (a,b) depending on the values of u at a and b. Now this means that

$$\mathcal{P}_e\left(\mathcal{P}_e u\right)(x) = \frac{\mathcal{P}_e u(x) - \mathcal{P}_e u(a)}{\mathcal{P}_e u(b) - \mathcal{P}_e u(a)} = \frac{\mathcal{P}_e u(x) - 0}{1 - 0} = \mathcal{P}_e u(x) \ ,$$

□

We now show that this 1D map can be applied to the edges of an element τ in a similar fashion to Example 4.1.

Definition 4.7. *Given a triangular element τ with nodes $\{n_i\}_{i=1}^3$ and edges $\{e_j\}_{j=1}^3$, where $e_j = \{n_j + t(n_{j+1} - n_j) \,|\, t \in [0,1]\}$ and $n_4 = n_1$, let $\mathcal{P}_{i,\tau} : C(\partial\tau) \to \mathbb{R}$ be defined by*

$$\mathcal{P}_{i,\tau} u \,|_{e_j} = (\delta(n_i, n_{j+1}) - \delta(n_i, n_j))\, \mathcal{P}_{e_j} u + \delta(n_i, n_j) \ , \tag{4.5}$$

for $i, j = 1, 2, 3$ and where

$$\delta(n_i, n_j) = \begin{cases} 1 & \text{if } n_i = n_j \\ 0 & \text{otherwise} \end{cases} \ .$$

Descriptively this means that the set of functionals $\{\mathcal{P}_{i,\tau}\}_{i=1}^3$ project the solution on to a nodal basis that preserves the fine scale properties. What we will show later in Section 4.4 is that the set $\{\mathcal{P}_{i,\tau} u\}_{i=1}^3$ forms a partition of unity (see Definition 2.28).

Remark 4.8. *We observe that if the function $u = 0$ then the boundary conditions obtained from $\{\mathcal{P}_{i,\tau} u\}_{i=1}^3$ are the linear functions such that $\mathcal{P}_{i,\tau} u(n_j) = \delta_{ij}$ for the nodes $\{n_j\}_{j=1}^3$ of the triangle τ.*

This previous remark is important because this incorporates the conventional process of oversampling into the adaptive framework. The oversampling method as defined in Section 4.5.1 is a one step method whereby a local problem is solved on the extended domain but only with linear boundary conditions and then these are combined to give multiscale basis functions on the element. It provides a good starting approximation to the solution when the approximation is updated iteratively by the ALG-MsFEM. Normally in the ALG-MsFEM algorithm, the oversampling method is stated as a separate step, here we will describe it as part of the full algorithm because of the result in Remark 4.8.

4.3.3 The local homogeneous problem

Now that we have the domain for the local problem defined by $\tilde{\tau}$ and the boundary conditions from $\{\mathcal{P}_{i,\tilde{\tau}} u\}_{i=1}^3$ then we can state the local homogeneous problem. Given a domain $\sigma \subset \Omega$ and boundary conditions ψ on $\partial\sigma$, find $\phi \in H^1(\sigma)$ with $\phi = \psi$ on $\partial\sigma$, such that

$$A_\sigma(\phi, v) = 0 \qquad \text{for all } v \in H_0^1(\sigma) \ . \tag{4.6}$$

In practice σ will be τ or $\tilde{\tau}$. The local problem (4.6) can be solved by any suitable means, for our implementation we chose to approximate using FEM on a fine mesh $\mathcal{T}_h(\sigma)$ on the domain σ. The accuracy of this local solve has implications for the accuracy across the whole domain but for now we assume h is sufficiently small not to produce a dominant error.

4.3.4 Finding the multiscale basis functions

To recap, the iterative cycle so far consists of finding an extended domain $\tilde{\tau}$ and corresponding boundary conditions $\{\mathcal{P}_{i,\tilde{\tau}}u\}_{i=1}^3$. The next step of the iterative cycle is to solve the local homogeneous problems on $\tilde{\tau}$ using these boundary conditions to get three oversampled basis functions $\left\{\Psi_{i,\tilde{\tau}}^{\mathrm{MS}}\right\}_{i=1}^3$. What we then have to do is define the multiscale basis functions $\left\{\Phi_{j,\tau}^{\mathrm{MS}}\right\}_{j=1}^3$ on the original element τ as a linear combination of the $\Psi_{i,\tilde{\tau}}^{\mathrm{MS}}$ such that the $\Phi_{j,\tau}^{\mathrm{MS}}$ are nodal. This means that the $\Phi_{j,\tau}^{\mathrm{MS}}$ take the form

$$\Phi_{j,\tau}^{\mathrm{MS}}(x) = \sum_{i=1}^{3} c_{ji}\Psi_{i,\tilde{\tau}}^{\mathrm{MS}}(x) \qquad \text{for } j = 1,2,3\ , \tag{4.7}$$

where

$$\Phi_{j,\tau}^{\mathrm{MS}}(n_k) = \delta_{jk} \qquad \text{for } j,k = 1,2,3\ , \tag{4.8}$$

and $\{n_k\}_{k=1}^3$ are the vertices of τ. The constants c_{ji} can be found by solving the linear system

$$C\Psi = I_3\ , \tag{4.9}$$

where $C_{ji} = c_{ji}$ and $\Psi_{ik}^{\mathrm{MS}} = \Psi_{\tilde{\tau},i}^{\mathrm{MS}}(n_k)$. We will show later in Section 4.4 that this definition allows us to preserve the partition of unity property (see Definition 2.28) in $\Psi_{i,\tilde{\tau}}$ for the $\Phi_{j,\tau}$, which is important to get good approximability for a Galerkin method.

We remark that if $\tilde{\tau} = \tau$ then $\Phi_{i,\tau}^{\mathrm{MS}} = \Psi_{i,\tilde{\tau}}^{\mathrm{MS}}$ for $i = 1,2,3$, which becomes relevant when we define the EDG1 ALG-MsFEM as a simplification of the general adaptive multiscale framework.

This completes the description of the iterative cycle. In Algorithm 1 below we summarize the process. Note that this iterative cycle is performed repeatedly as part of a larger algorithm outlined in Algorithm 2 in Section 4.5.

Algorithm 1 The iterative step

1: Given an initial solution u, an element τ and a corresponding extended element $\tilde{\tau}$:
2: Find $\left\{\Psi_{\tau,j}^{\mathrm{MS}}\right\}_{j=1}^{3}$ on $\partial\tilde{\tau}$ by calculating $\Psi_{\tau,j}^{\mathrm{MS}}\Big|_{e_i} = \mathcal{P}_{j,e_i}u$ for $i,j=1,2,3$.
3: Solve the local homogenous problem (4.6) to get $\left\{\Psi_{\tau,j}^{\mathrm{MS}}\right\}_{j=1}^{3}$ on the interior of $\tilde{\tau}$.
4: Find c_{ij} so that $\Phi_{\tau,i} = \sum_{j=1}^{3} c_{ij}\Psi_{\tau,j}$ and $\Phi_{\tau,i}(n_k) = \delta_{ik}$ for the vertices $\{n_k\}_{k=1}^{3}$ of τ.
5: Calculate $\left\{\Phi_{\tau,i}^{\mathrm{MS}}\right\}_{i=1}^{3}$ using these c_{ij} and Ψ_j.
6: Pass the result to Algorithm 2

It is important to note that Algorithm 1 must be performed on each element τ that requires a multiscale basis function, and so it is only one stage in a larger iterative process. In particular Algorithm 1 yields the multiscale functions on each τ but there is still the important step of joining them together and then using the resulting basis functions to solve the global problem derived from the bilinear form in (5.3). To this end we give a formal definition for the global Φ_i^{MS} by joining the local multiscale basis functions $\Phi_{i,\tau}^{\mathrm{MS}}$ that are non-zero at the global node n_i of the mesh $\mathcal{T}_H(\Omega)$ together, this allows us to move between a local and global setting for the basis functions.

Definition 4.9. *For any node $n_i \in \mathcal{N}(\mathcal{T}_H(\Omega))$ and element $\tau \in \mathcal{T}_H(\Omega)$ let Φ_i^{MS} be defined on all of Ω by*

$$\Phi_i^{MS}|_\tau = \Phi_{i,\tau}^{MS} \ . \tag{4.10}$$

This now gives a finite dimensional approximation space

$$V_H = \operatorname{span}\left\{\Phi_i^{\mathrm{MS}}\right\}_{n_i\in\mathcal{N}(\Omega)} \ , \tag{4.11}$$

(where $V_H \subset H^1(\Omega)$ if the basis functions are continuous across element edges), moreover if we take only the interior nodes then

$$V_{H,0} = \operatorname{span}\left\{\Phi_i^{\mathrm{MS}}\right\}_{n_i\in\mathcal{N}_0(\Omega)} \ , \tag{4.12}$$

is a set of test functions for solving a finite dimensional version of Problem 5.3 (similarly $V_{H,0} \subset H_0^1(\Omega)$ if the basis functions are continuous across element edges). The global problem (4.13) is stated below.

Problem 4.10 (Global Problem)**.** *Let $\{\Phi_i\}_{x_i\in\mathcal{N}_0(\mathcal{T}_H(\Omega))}$ be a finite set of basis func-*

tions. Find $u_H \in V_H$ *with* $u_H = g$ *at the nodes on* $\partial\Omega$*, such that*

$$\sum_{x_i \in \mathcal{N}_0(\mathcal{T}_H(\Omega))} U_i A\left(\Phi_i, \Phi_j\right) = F\left(\Phi_j\right) \quad - \sum_{x_k \in \mathcal{N}_D(\mathcal{T}_H(\Omega))} g(x_k) A\left(\Phi_k, \Phi_j\right) \tag{4.13}$$

for all $x_j \in \mathcal{N}_0(\mathcal{T}_H(\Omega))$*, and*

$$u_H = \sum_{x_i \in \mathcal{N}_0(\mathcal{T}_H(\Omega))} U_i \Phi_i + \sum_{x_k \in \mathcal{N}_D(\mathcal{T}_H(\Omega))} g(x_k) \Phi_k \ . \tag{4.14}$$

This solution is found by solving the system

$$K\mathbf{U} = \mathbf{F} - K^D \mathbf{g} \tag{4.15}$$

where $K_{ij} = A(\Phi_i, \Phi_j)$*,* $F_j = \mathbf{F}(\Phi_j)$ *for* $x_i, x_j \in \mathcal{N}_0(\mathcal{T}_H(\Omega))$ *and* $K_{ij}^D = A(\Phi_i, \Phi_j)$*,* $\mathbf{g}_i = g(x_i)$ *for* $x_i \in \mathcal{N}_D(\mathcal{T}_H(\Omega))$*,* $x_j \in \mathcal{N}_0(\mathcal{T}_H(\Omega))$*.*

Using this newly obtained approximation u_H, we can apply Algorithm 1 again to each element and repeat the process iteratively. Now that each component is in place we give the whole algorithm for the adaptive multiscale method framework in the following section and examine how slight alterations to certain steps results in the various algorithms in [36] as well as a new enhanced version of their method. A flow diagram of the iterative concept is given below.

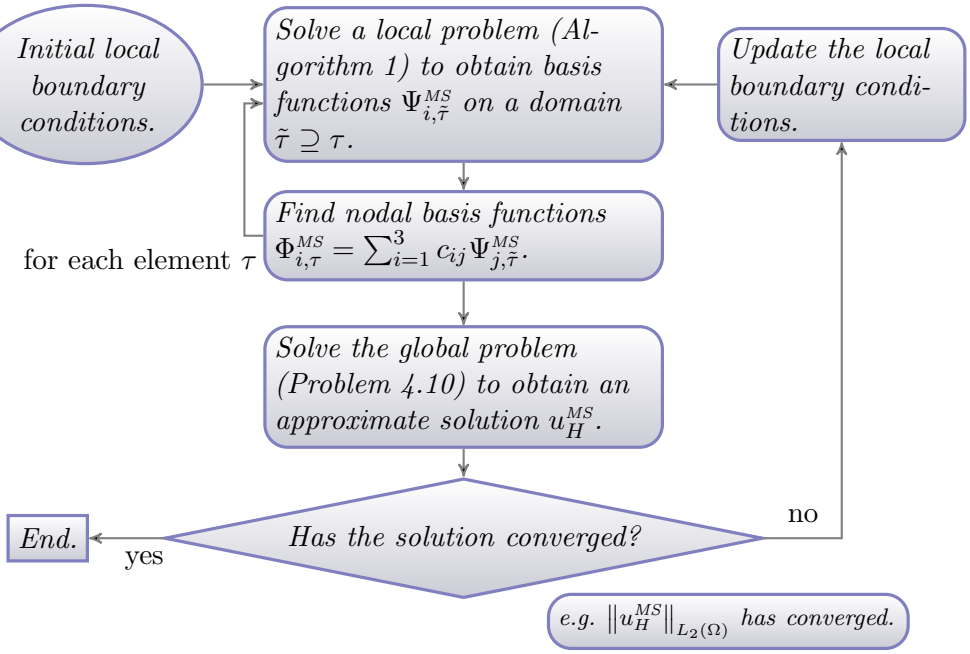


Figure 4-2: *Flowchart for the basis function iteration concept.*

4.4 Properties of the adaptive multiscale method

In the previous section the process by which the adaptive multiscale basis functions are found was examined, however before stating the algorithm that forms the adaptive multiscale method we first state and prove some of the key properties of the method. The first property is that the basis functions $\Phi^{\mathrm{MS}}_{\tau,i}$ are obviously nodal from (4.8), $\Phi^{\mathrm{MS}}_{\tau,i}(n_j) = \delta_{ij}$ for the nodes n_j of τ where $i, j = 1, 2, 3$.

The next step is to show that the multiscale basis functions $\Phi_{\tau,i}$ form a partition of unity (see Definition 2.28) on τ. This is done in several stages by first showing that the edge mapping $\{\mathcal{P}_{\tilde{\tau},i}u\}$ forms a partition of unity on the boundary of the extended triangle $\tilde{\tau}$. Using the uniqueness of the solution on the interior of the domain we are then able to show that the process of finding the basis functions on τ preserves the partition of unity property.

Lemma 4.11. *For any $u \in C(\partial\tilde{\tau})$, $\{\mathcal{P}_{i,\tilde{\tau}}u\}_{i=1}^{3}$ forms a partition of unity of $\partial\tilde{\tau}$.*

Proof. Denote the edges of $\tilde{\tau}$ by $\tilde{e}_k$ for $k = 1, 2, 3$ and the nodes of $\tilde{\tau}$ by n_j for $j = 1, 2, 3$. Then (4.5) implies

$$\mathcal{P}_{i,\tilde{\tau}}u\big|_{\tilde{e}_k} = \begin{cases} 1 - \mathcal{P}_{\tilde{e}_k}u & \text{if } \; i = k \\ \mathcal{P}_{\tilde{e}_k}u & \text{if } \; i = k+1 \\ 0 & \text{otherwise} \end{cases}$$

where $n_4 := n_1$. This is non-zero only when $i = k, k+1$. Therefore

$$\sum_{i=1}^{3} \mathcal{P}_{i,\tilde{\tau}}u\big|_{\tilde{e}_k} = 1 - \mathcal{P}_{\tilde{e}_k}u + \mathcal{P}_{\tilde{e}_k}u = 1$$

for any $k = 1, 2, 3$. Therefore $\sum_{i=1}^{3} \mathcal{P}_{i,\tilde{\tau}}u = 1$ on $\partial\tilde{\tau}$. □

Now that we have shown that the functions $\mathcal{P}_{i,\tilde{\tau}}u$ form a partition of unity on $\tilde{\tau}$ we show that this property extends to the interior of $\tilde{\tau}$ as well.

Lemma 4.12. *The functions $\{\Psi_{i,\tilde{\tau}}\}_{i=1}^{3}$ form a partition of unity on $\tilde{\tau}$.*

Proof. The basis functions $\Psi_{i,\tilde{\tau}}$ on $\tilde{\tau}$ solve the local homogeneous problem (4.6) with

boundary condition $\mathcal{P}_{i,\tilde{\tau}}u$ (see Section 4.3.4). Therefore

$$A\left(\sum_{i=1}^{3}\Psi_{i,\tilde{\tau}},v\right)=\sum_{i=1}^{3}A\left(\Psi_{i,\tilde{\tau}},v\right)=0 \qquad \text{for any } \; v\in H_0^1(\tilde{\tau}) \;,$$
$$\sum_{i=1}^{3}\Psi_{i,\tilde{\tau}}=\sum_{i=1}^{3}\mathcal{P}_{i,\tilde{\tau}}u=1 \qquad \text{on } \; \partial\tilde{\tau} \;, \tag{4.16}$$

by Lemma 4.11. Note that $\Phi=1$ satisfies $A(\Phi,v)=0$ for any $v\in H_0^1(\tilde{\tau})$ and $\Phi=1$ on $\partial\tilde{\tau}$. Since the solution to the local problem (4.16) is unique then $\sum_{i=1}^{3}\Psi_{i,\tilde{\tau}}=1$ on τ. □

Using the previous two lemmas we show that the basis functions $\left\{\Phi_{i,\tau}^{\mathrm{MS}}\right\}_{i=1}^{3}$, which are a linear combination of $\left\{\Psi_{i,\tilde{\tau}}^{\mathrm{MS}}\right\}_{i=1}^{3}$ (see Section 4.3.4), inherit the partition of unity property on τ.

Proposition 4.13. *The set of functions* $\left\{\Phi_{i,\tau}^{MS}\right\}_{i=1}^{3}$ *forms a partition of unity on* τ.

Proof. The basis functions $\Phi_{i,\tau}^{\mathrm{MS}}$ are a linear combination of $\Psi_{i,\tilde{\tau}}^{\mathrm{MS}}$ from (4.7) where the coefficients c_{ij} are found by solving the matrix system (4.9). Therefore

$$\delta_{ij}=\Phi_{i,\tau}^{\mathrm{MS}}(n_j)=\sum_{k=1}^{3}c_{ik}\Psi_{k,\tilde{\tau}}^{\mathrm{MS}}(n_j) \qquad \text{for } i,j=1,2,3 \;, \tag{4.17}$$

where n_j are the nodes of τ. In (4.9) the previous equation was abbreviated to $C\Psi=I_3$. Note this implies

$$\Psi C=(\Psi C)\,\Psi\Psi^{-1}=\Psi\left(C\Psi\right)\Psi^{-1}=\Psi\Psi^{-1}=I_3 \;,$$

and hence

$$\sum_{k=1}^{3}\Psi_{i,\tilde{\tau}}^{\mathrm{MS}}(n_k)c_{kj}=\delta_{ij} \qquad \text{for } \; i,j=1,2,3 \;.$$

Then since $\sum_{i=1}^{3}\Psi_{i,\tilde{\tau}}(x)=1$ for any $x\in\tilde{\tau}$ by Lemma 4.12,

$$\sum_{k=1}^{3}c_{kj}=\sum_{k=1}^{3}1\cdot c_{kj}=\sum_{k=1}^{3}\left(\sum_{i=1}^{3}\Psi_{i,\tilde{\tau}}(n_k)\right)c_{kj}=\sum_{i=1}^{3}\sum_{k=1}^{3}\Psi_{i,\tilde{\tau}}(n_k)c_{kj}=\sum_{i=1}^{3}\delta_{ij}=1 \tag{4.18}$$

Using (4.7) we finally obtain

$$\sum_{i=1}^{3} \Phi_{i,\tau}^{\mathrm{MS}} = \sum_{i=1}^{3} \sum_{j=1}^{3} c_{ij} \Psi_{j,\tilde{\tau}}^{\mathrm{MS}} = \sum_{j=1}^{3} \left(\sum_{i=1}^{3} c_{ij} \right) \Psi_{j,\tilde{\tau}}^{\mathrm{MS}} = \sum_{j=1}^{3} \Psi_{j,\tilde{\tau}}^{\mathrm{MS}} = 1 \ ,$$

by (4.18) and Lemma 4.12. □

4.5 Variants of adaptive multiscale methods

In this section we give a general description of the adaptive multiscale finite element framework. Algorithm 2 is adapted from [36] where it was formulated for the finite volume method. Algorithm 2 describes both the EDG1 and EDG2 Adaptive Local Global Multiscale Finite Element Method (ALG-MsFEM) proposed in [36] as well as an enhanced ALG-MsFEM, which we used here.

The major advantage of the local global methods is that global problems are only solved on a coarse grid $\mathcal{T}_H$ where H is much larger than the fine grid h used for the local problems. This allows a solution to be found when it is unfeasible to solve the global problem on the fine mesh. It is also very useful if many problems with the same coefficient $\mathcal{A}$ but different boundary conditions and source terms are to be solved (p39 [37], [49]) by storing the basis functions from a previous calculation with a specific source and simply reusing them in the coarse global problem.

4.5.1 The oversampled method

All of the adaptive local-global methods start with an initial step of an oversampled multiscale finite element method. The oversampled FEM uses Algorithm 1 as described in Section 4.3 solving local problems on an extended element but with linear boundary conditions. The linear Dirichlet conditions are a consequence of using an initial solution of $u = 0$ in the projection $\mathcal{P}_{i,\tau}$. The oversampled FEM follows Algorithm 2 but with only one cycle, it is not an iterative process.

Once the initial basis functions have been found using the linear combination from the extended element (4.9) then they may be discontinuous across the edges. For the EDG1 and EDG2 ALG-MsFEM's this results in a discontinuous solution which then needs to be averaged across the edges of the mesh $\mathcal{T}_H(\Omega)$. The resulting basis functions are still non-conforming but provide a good initial approximation to the multiscale basis functions.

Algorithm 2 The adaptive multiscale method framework

Set $u_H = 0$ initially.
repeat
  for each element $\tau \in \mathcal{T}_H(\Omega)$ **do**
    if [EDG1 ALG-MsFEM] and iteration number> 1 **then**
      Set $\tilde{\tau} = \tau$.
    else
      Find an extended domain $\tilde{\tau}$ around τ.
    end if
    Use the iterative step (Algorithm 1) with u_H, τ and $\tilde{\tau}$ to find $\left\{\Phi_{\tau,i}^{\mathrm{MS}}\right\}_{i=1}^{3}$.
  end for
  if [enhanced ALG-MsFEM] **then**
    Average the edges of the $\Phi_{\tau,i}^{\mathrm{MS}}$ with their neighbours.
    Re-solve the local problem (4.6) on each element τ.
  end if
  For each node $n_i \in \mathcal{T}_H(\Omega)$ set $\Phi_i^{\mathrm{MS}}| = \Phi_{\tau,i}^{\mathrm{MS}}$.
  Using this basis of $\{\Phi_i^{\mathrm{MS}}\}$ solve the global problem (Problem 4.10) to find a new u_H.
  if [EDG2 ALG-MsFEM] or [Oversampled FEM] **then**
    Average u_H on the edges of the elements in $\mathcal{T}_H(\Omega)$.
  end if
until u_H has converged or using [Oversampled FEM].

4.5.2 The EDG1 ALG-MsFEM

The EDG1 ALG-MsFEM performs the oversampled method first to obtain an initial approximation to the solution. Using this initial solution and the iterative cycle (Algorithm 1) gives a new set of boundary conditions to define a new set of basis functions $\left\{\Phi_{i,\tau}^{\mathrm{MS}}\right\}$.

The main feature of the EDG1 method is that the extended domain $\tilde{\tau}$ is set as τ. This means that if two neighbouring elements τ_1, τ_2 share an edge e then the edge projection $\mathcal{P}_e$ is the same for each element, i.e.

$$\mathcal{P}_{i,\tau_1} u \mid_e = \mathcal{P}_{i,\tau_2} u \mid_e \ .$$

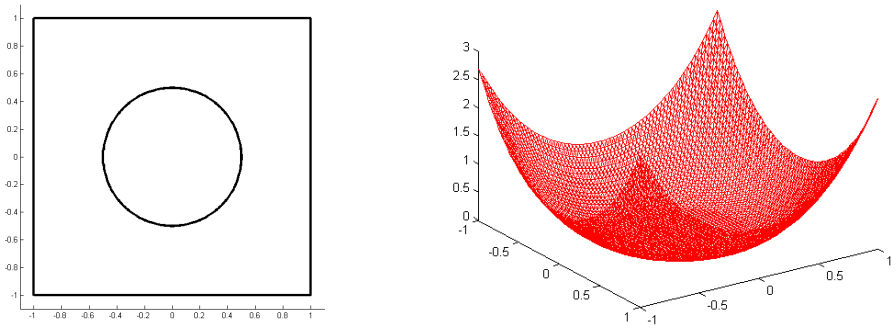
This results in basis functions that are continuous at the element edges and hence the method is conforming. The fact that there is no extended region for the local solve is also the method's main drawback. Since the local solve is on the element itself there is no mechanism for transporting information across the domain and hence the bad

boundary condition from the oversampled method can not be improved. Consequently the rate of convergence compared to the standard FEM is not improved as demonstrated by the numerical example below, however the error over the standard finite element method is improved. Note also that the local boundary conditions do not update after the second iteration, so rather than an iterative method the EDG1 ALG-MsFEM can be considered more as a two-step method.

We now give a numerical example to show the convergence rate of the EDG1 ALG-MsFEM. The example comes from [27] and is discussed in more detail in Section 4.6.1. Here $\Omega = [-1, 1]^2$ with a single circular inclusion, Figure 4-3(a), such that

$$\mathcal{A}(x) = \begin{cases} \mathcal{A}_1 & r < r_0 \\ \mathcal{A}_0 & r \geq r_0 \end{cases}, \tag{4.19}$$

where $r = \left(x^2 + y^2\right)^{\frac{1}{2}}$ and $r_0 = \pi/6.28$ so as not to be resolved by any uniform mesh.



(a) (b)

Figure 4-3: *The domain of this experiment with a single circular inclusion (left) and an example exact solution where $\mathcal{A}_1 = 10^5$ (right).*

The details of the problem are given in Section 4.6.1 but it is designed to have an exact solution given by

$$u(r, \theta) = \begin{cases} \frac{r^3}{\mathcal{A}_1} & r < r_0 \\ \frac{r^3}{\mathcal{A}_0} + \left(\frac{1}{\mathcal{A}_1} - \frac{1}{\mathcal{A}_0}\right) r_0^3 & r \geq r_0 \end{cases}. \tag{4.20}$$

Since u is known analytically we may compute the L_2 error, $\|u - u_H^{\text{MS}}\|_{L_2(\Omega)}$. These results are stated below in Table 4.1

The results for the $\mathcal{A}_1 \to \infty$ case appear to have better convergence rates than the $\mathcal{A}_0 \to \infty$ case but it is actually only at the start for the first few values of H and then

H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	6.7562E-02	6.7816E-02	6.8331E-02	6.8201E-02	6.7277E-02
1/8	1.7076E-02	3.2327E-02	2.8592E-02	3.0541E-02	3.6709E-02
1/16	4.1988E-03	7.9879E-03	1.0703E-02	1.3253E-02	1.7341E-02
1/32	1.4441E-03	2.7267E-03	4.2114E-03	6.5874E-03	9.0844E-03
1/64	8.5095E-04	1.3276E-03	2.6162E-03	4.2850E-03	5.4319E-03
Rate	1.6186	1.4917	1.2177	1.0198	0.92758

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	1.0973E-02	1.6500E-02	3.8042E-02	5.6285E-02	6.8416E-02
1/8	4.5724E-03	1.4952E-02	2.8064E-02	4.3161E-02	4.8883E-02
1/16	2.2424E-03	5.7076E-03	1.2877E-02	2.2280E-02	2.6350E-02
1/32	1.4859E-03	3.0733E-03	6.3363E-03	1.0686E-02	1.3300E-02
1/64	5.6361E-04	1.0111E-03	2.0601E-03	3.7520E-03	4.8338E-03
Rate	1.0188	1.0340	1.0561	0.9828	0.9524

Table 4.1: *L_2 norm of the error using EDG1 ALG-MsFEM where $\mathcal{A}_0 = 1$ and $\mathcal{A}_1 \to \infty$ (top) and $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ (bottom) with $M = 32$.*

it levels out to $O(H)$. These results simply act as a demonstration to show that while they are smaller than the standard finite element error (Section 4.6.1 Table 4.3) they do not improve the rate of convergence and it is still $O(H)$. Recall that if the coefficient were smooth we would expect an optimal convergence rate in the L_2 norm of $O(H^2)$. The local problems for the basis functions (Step 3 in Algorithm 1) were done using the Immersed FEM [62] with a uniform fine grid with diameter $h = (1/M)H$.

4.5.3 The EDG2 ALG-MsFEM

The EDG2 ALG-MsFEM has exactly the same form as the oversampled method in Section 4.5.1 but rather than using the linear boundary conditions it uses the full edge projection $\mathcal{P}_{i,\tau}u$ where u is the current guess and is used repeatedly in an iterative process.

The problem with this method is that it too produces basis functions that are discontinuous across the elements of the mesh, thus producing a discontinuous solution. This method again simply averages the discontinuous solution across the element edges to make it continuous. The use of information on the extended domain does improve the convergence rate for the problem in (4.19) and (4.20), the results are given in Table 4.2.

In comparison to Table 4.1, the results in Table 4.2 are much better and we see that the order of convergence has improved to $O(H^2)$ for the L_2 norm. It still however has the disadvantage that the basis functions are discontinuous. It is worth noting also that the EDG2 method failed to converge to a solution in all cases. Instead the tests

H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	8.8521E-02	9.6363E-02	9.9681E-02	9.5768E-02	9.4430E-02
1/8	2.1025E-02	2.4181E-02	2.2030E-02	2.0920E-02	2.0740E-02
1/16	5.2057E-03	5.3464E-03	5.7211E-03	5.0941E-03	4.9456E-03
1/32	1.2291E-03	1.2980E-03	1.3667E-03	1.2948E-03	1.2171E-03
1/64	3.3083E-04	3.5493E-04	4.0325E-04	4.0336E-04	3.0668e-04
Rate	2.0224	2.0389	1.9910	1.9797	2.0624

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	1.5371E-02	1.5026E-02	1.4998E-02	1.4924E-02	1.4916E-02
1/8	3.8948E-03	2.7756E-03	3.2367E-03	3.5289E-03	3.6045E-03
1/16	1.1889E-03	8.9006E-04	9.3043E-04	1.0056E-03	1.0060E-03
1/32	3.0762E-04	2.8036E-04	2.9540E-04	2.8032E-04	3.0061E-04
1/64	7.2648E-05	7.0962E-05	7.1425E-05	8.6866E-05	8.4060E-05
Rate	1.9112	1.8760	1.8882	1.8503	1.8526

Table 4.2: *L_2 norm of the error using EDG2 ALG-MsFEM where $\mathcal{A}_0 = 1$ and $\mathcal{A}_1 \to \infty$ (top) and $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ (bottom) with $M = 32$.*

were terminated after 10 iterations. The solution appeared to oscillate with errors about $\pm 3\%$ of the final value after 10 iterations. This problem does not happen with the enhanced ALG-MsFEM described next.

4.5.4 The enhanced ALG-MsFEM

The EDG2 ALG-MsFEM proposed by Durlofsky, Efendiev and Ginting [36] post processes the approximate solution by averaging along the edges to create a continuous solution. What we propose here is to introduce an enhanced version of the EDG2 ALG-MsFEM that makes it conforming. An additional consequence of the conforming method is that the enhanced ALG-MsFEM removes the need to post-process the approximate solution by averaging the values along the edges to produce a continuous solution.

The alteration to the framework for this method is to introduce another stage after the iterative step has been performed for each element. Once an initial discontinuous multiscale basis function has been found we average the values of this basis function across edges and then re-solve the local homogeneous problem (4.6) only on τ and not $\tilde{\tau}$. This is different from the method proposed in [36] because it averages the basis function edges rather than the approximate solution. Therefore the enhanced ALG-MsFEM starts off with a conforming basis for the global problem (Problem 4.10) rather than making the solution continuous after the global solve is performed. The cost is the solution of additional local problems after the basis function is averaged. This method actually produces a slightly smaller error than the EDG2 ALG-MsFEM as we shall see

in the numerical examples in Section 4.6.1.

The use of the extended domain allows the enhanced ALG-MsFEM to transport information across the domain just as the EDG2 ALG-MsFEM does but by averaging the edges of the basis functions we automatically get a continuous solution. As in Section 3.1.6, consider two neighbouring elements τ_1, τ_2 in $\mathcal{T}_H(\Omega)$ that share an edge e, set

$$\Phi_i^{\mathrm{MS}} \mid_e = \frac{\Phi_{i,\tau_1}^{\mathrm{MS}} + \Phi_{i,\tau_2}^{\mathrm{MS}}}{2} ,$$

and then re-solve the local homogeneous problem with these new boundary conditions. The averaging process increases the support to a star shape as in Figure 4-4, this is because the outer elements now no longer have zero value on their boundaries that link to the original support. This makes the basis functions non-zero in the additional support regions. This increase in the support of the basis functions increases the number of non-zeros in the stiffness matrix but the support is still relatively small meaning that the matrix is still very sparse, consequently solve times are not impacted significantly.

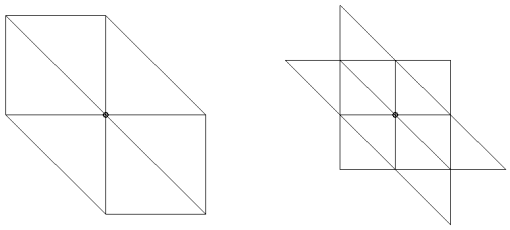


Figure 4-4: *An example of how the edge averaging can increase the support of the basis functions, the original on the left and extended version on the right.*

4.6 Numerical convergence analysis and properties

To show the power of the enhanced ALG-MsFEM we examine several classes of examples. They all demonstrate how the method gives a superior convergence rate compared to the standard FEM. In all cases the domain is taken to be $\Omega = [-1, 1] \times [-1, 1]$, and the problem is: Find $u \in H^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \mathcal{A} \nabla v = \int_\Omega f v \qquad \text{for any } v \in H_0^1(\Omega) ,$$
$$u = g \quad \text{on } \Gamma_D , \qquad \frac{\partial u}{\partial n} = 0 , \quad \text{on } \Gamma_N , \tag{4.21}$$

where f, g, the Dirichlet boundary Γ_D and the Neumann boundary Γ_N are problem specific. In Sections 4.6.1 to 4.6.4 we will consider interface problems with coefficient

$$\mathcal{A}(x) = \begin{cases} \mathcal{A}_1 & x \in \Omega_1 \\ \mathcal{A}_0 & x \in \Omega_0 \end{cases} ,$$

where Ω_0 and Ω_1 are problem specific. In the final set of simulations (Section 4.6.5) we consider $\mathcal{A}(x)$ as a representation of a certain log-normal random field.

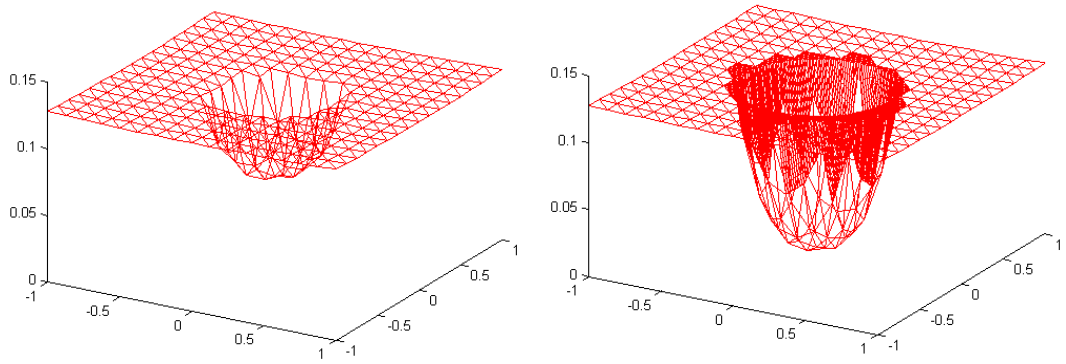
4.6.1 High contrast examples

This first example comes from [27] and the purpose of repeating it here is to validate the ALG-MsFEM and show that it performs as well as the highly specific robust MsFEM used there. In fact the L_2 errors in approximation are slightly smaller than those in [27]. The problem we are solving uses $\Omega_1 = \{x \in \Omega \,|\, r < r_0\}$ and $\Omega_0 = \Omega \backslash \Omega_1$ as in (4.19). This experiment is unusual in that it was designed so that it has an exact solution given by (4.20), which leads to $\Gamma_D = \partial\Omega$, $\Gamma_N = \emptyset$ and

$$f = -9r \ , \qquad\qquad g := \frac{r^3}{\mathcal{A}_0} + \left(\frac{1}{\mathcal{A}_1} - \frac{1}{\mathcal{A}_0}\right) r_0^3 \ .$$

in (4.21). We ran both the standard finite element method and enhanced ALG-MsFEM with uniform meshes over Ω with element size H varying from $\frac{1}{4}$ down to $\frac{1}{64}$. Figure 4-5 shows the difference between the two solutions for a specific H. The multiscale basis functions allow far more detail and by measuring the error in approximation in the L_2 norm we can see how it is also much more accurate. In this specific example the multiscale basis functions capture the jump in the gradient inside the coarse elements much better and thus the approximate solution is much more accurate in the Adaptive MsFEM case because the bowl shape is deeper like the true solution, in fact the minimum should be zero. The numerical accuracy is considered in the tables below.

Table 4.3 gives the L_2 error for the standard finite element method, $\|u - u_H\|_{L_2(\Omega)}$, in both the $\mathcal{A}_1 \to \infty$ case (top) and the $\mathcal{A}_0 \to \infty$ case (bottom). Table 4.4 describes similar results but for the enhanced ALG-MsFEM error, $\|u - u_H^{\text{MS}}\|_{L_2(\Omega)}$. The numerical results show what is expected from the error bound stated in Theorem 2.60 in Chapter 2, that the standard finite element error in the L_2 norm $\|u - u_H\|_{L_2(\Omega)}$ is only $O(H)$. Here it is important to note however that the error does not depend on the contrast as many other results have stated in the past (see Section 1.2.1) but is contrast independent as predicted by Theroem 2.60. This is true for both the standard FEM and enhanced ALG-MsFEM.



(a) Standard FEM (b) Enhanced ALG-MsFEM

Figure 4-5: *Plot showing the approximate solution for both the standard FEM and adaptive MsFEM for the case when* $\mathcal{A}_0 = 10^3$ *and* $H = \frac{1}{8}$*. The adaptive MsFEM uses a subgrid on cut elements with* $h = \frac{1}{64}$*.*

H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	9.0690e-02	1.1390e-01	1.5289e-01	2.2378e-01	2.7508e-01
1/8	3.2707e-02	5.0821e-02	6.4377e-02	7.9325e-02	8.9021e-02
1/16	1.4064e-02	2.4362e-02	3.0716e-02	3.4457e-02	3.5588e-02
1/32	6.8547e-03	1.2323e-02	1.4760e-02	1.5506e-02	1.5617e-02
1/64	3.3615e-03	6.0662e-03	7.3507e-03	7.6623e-03	7.7762e-03
Rate	1.1762	1.0506	1.0882	1.2091	1.2800

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	3.8417e-02	6.4700e-02	7.1673e-02	7.2556e-02	7.2646e-02
1/8	2.3114e-02	4.3309e-02	5.0189e-02	5.1152e-02	5.1253e-02
1/16	1.2233e-02	2.2410e-02	2.8010e-02	3.0120e-02	3.0450e-02
1/32	6.6788e-03	1.2566e-02	1.5077e-02	1.5656e-02	1.5729e-02
1/64	3.3645e-03	6.0060e-03	6.8098e-03	6.9357e-03	6.9493e-03
Rate	0.8818	0.8644	0.8526	0.8482	0.8476

Table 4.3: *L_2 norm of the error using the standard finite element method where* $\mathcal{A}_0 = 1$ *and* $\mathcal{A}_1 \to \infty$ *(top) and* $\mathcal{A}_0 \to \infty$ *and* $\mathcal{A}_1 = 1$ *(bottom)*

The next point to note is that the adaptive multiscale finite element error in the L_2 norm, $\|u - u_H^{\text{MS}}\|_{L_2(\Omega)}$, is $O(H^2)$ and also independent of contrast. The enhanced ALG-MsFEM has restored the rate of convergence as if there were no loss of regularity (All rates were found by linear regression), and to achieve the same threshold of error we need to solve a much smaller matrix system. For example when $\mathcal{A}_0 = 10^3$ then the standard FEM produces an error of 6.8098×10^{-3} for $H = \frac{1}{64}$, therefore the global stiffness matrix has $O(128^2)$ non-zero entries (this being the rate at which the matrix system is solved by a sparse solver). In contrast the enhanced ALG-MsFEM obtains an error of 3.2576×10^{-3} for $H = \frac{1}{8}$, meaning the stiffness matrix only has $O(8^2)$ non-zero entries. While there are local solves to be done, these can all be done in parallel making

H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	6.9540e-02	6.8936e-02	6.8305e-02	6.7979e-02	6.7816e-02
1/8	1.7280e-02	1.7272e-02	1.7159e-02	1.6911e-02	1.6796e-02
1/16	4.3736e-03	4.3683e-03	4.3275e-03	4.2114e-03	4.1397e-03
1/32	1.0984e-03	1.0984e-03	1.0854e-03	1.0446e-03	1.0271e-03
1/64	2.7547e-04	2.7527e-04	2.7149e-04	2.5976e-04	2.6981e-04
Rate	1.9935	1.9911	1.9933	2.0081	1.9979

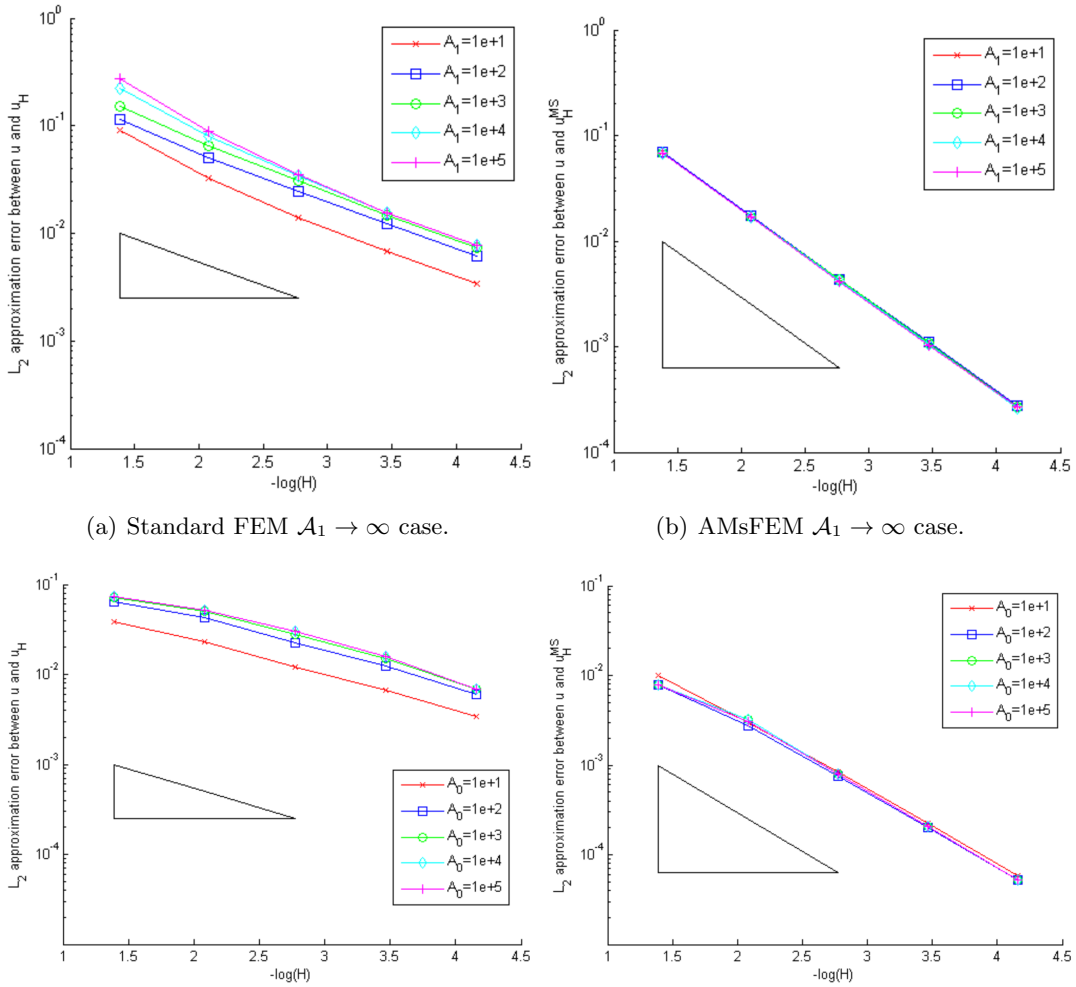
H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	1.0035e-02	7.9146e-03	7.7646e-03	7.7677e-03	7.8678e-03
1/8	2.9564e-03	2.7738e-03	3.2576e-03	3.2334e-03	3.0956e-03
1/16	8.4668e-04	7.5851e-04	7.8950e-04	8.0608e-04	8.0385e-04
1/32	2.2491e-04	2.0206e-04	2.0435e-04	2.0476e-04	2.0437e-04
1/64	5.8141e-05	5.2423e-05	5.2443e-05	5.2476e-05	5.2849e-05
Rate	1.8579	1.8255	1.8415	1.8400	1.8357

Table 4.4: *L_2 norm of the error using AMsFEM with an immersed FEM subgrid solve where $\mathcal{A}_0 = 1$ and $\mathcal{A}_1 \to \infty$ (top) and $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ (bottom) with $M = 32$.*

the global solve the main cost therefore even though this $O(8^2)$ matrix must be solved several times for the adaptive method, typically 4 or 5 iterations, it is still less complex than the $O(128^2)$ system.

This is a good point to discuss the impact of the choice of $M := H/h$. If the local problem is solved using a standard finite element method then M should be chosen greater than $1/H$ in order to ensure that optimal convergence is obtained independent of the contrast. Note that this significantly increases the complexity of the serial algorithm. The standard method requires $O(H^{-4})$ operations to achieve an $O(H^2)$ error in the L_2 norm while the multiscale method would require $O(H^{-2}M^2I)$ where I is the small number of iterations. The advantage comes when the multiscale algorithm is performed in parallel, then it only requires $O(H^{-2}M^2I/P)$ operations plus the overhead associated with communication. The primary focus of this algorithm is not to be faster than the standard finite element method but to provide a more accurate solution when there is an extreme value of the contrast as well as when singularities are present. In this situation the standard FEM requires $O(H^{-4\epsilon})$ when the solution is in $H^{1+\epsilon}(\Omega)$. However, the link between the fine mesh size and the contrast requires further study, as the rest of the examples in this chapter will show, it is not always necessary to have a fine h this small. The superior convergence is often observed when the subgrid mesh is comparatively coarse. Coarsening the subgrid mesh introduces a new consideration, when a coarse subgrid mesh is used the iterative process requires more iterations before convergence. Further study is also required to observe how the number of iterations increases with a coarser subgrid mesh.

The results in Tables 4.3 and 4.4 are displayed as log-plots in Figure 4-6. The triangles plotted help to give an indication of the $O(H)$ convergence for the standard FEM results (left column) and then the $O(H^2)$ convergence for the AMsFEM results, having gradients 1 and 2 respectively on a log plot.



(a) Standard FEM $\mathcal{A}_1 \to \infty$ case.

(b) AMsFEM $\mathcal{A}_1 \to \infty$ case.

(c) Standard FEM $\mathcal{A}_0 \to \infty$ case.

(d) AMsFEM $\mathcal{A}_0 \to \infty$ case.

Figure 4-6: *Log plots of the standard FEM L_2 errors, $\|u - u_H\|_{L_2(\Omega)}$, in Table 4.3 against $-log(H)$ (graphs 4-6(a) and 4-6(c)) as well as the adaptive MsFEM L_2 errors, $\|u - u_H^{MS}\|_{L_2(\Omega)}$, in Table 4.4 (graphs 4-6(b) and 4-6(d)).*

4.6.2 Multiple inclusions

We next consider the case of multiple inclusions. For this experiment we use the following definitions for the coefficient; let $c_1 = (0, -0.5)$, $c_2 = (0, 0.5)$ and $r_1 = r_2 = \pi/12.56$ then define

$$\Omega_1 = \{\|x - c_1\| < r_1\} \cup \{\|x - c_2\| < r_2\} \qquad \Omega_0 = \Omega \backslash \Omega_1 \ .$$

We however use the same load function f and the same boundary conditions g as the single inclusion problem previously:

$$f = -9r \qquad g = \frac{r^3}{\mathcal{A}_0} + \left(\frac{1}{\mathcal{A}_1} - \frac{1}{\mathcal{A}_0}\right) r_0^3 \ ,$$

where $r = \left(x^2 + y^2\right)^{\frac{1}{2}}$ and $r_0 = \pi/6.28$. The exact solution is unknown so to obtain a reference solution a very fine mesh for the AMsFEM was used, where $H = 1/128$ and $h = 1/4096$.

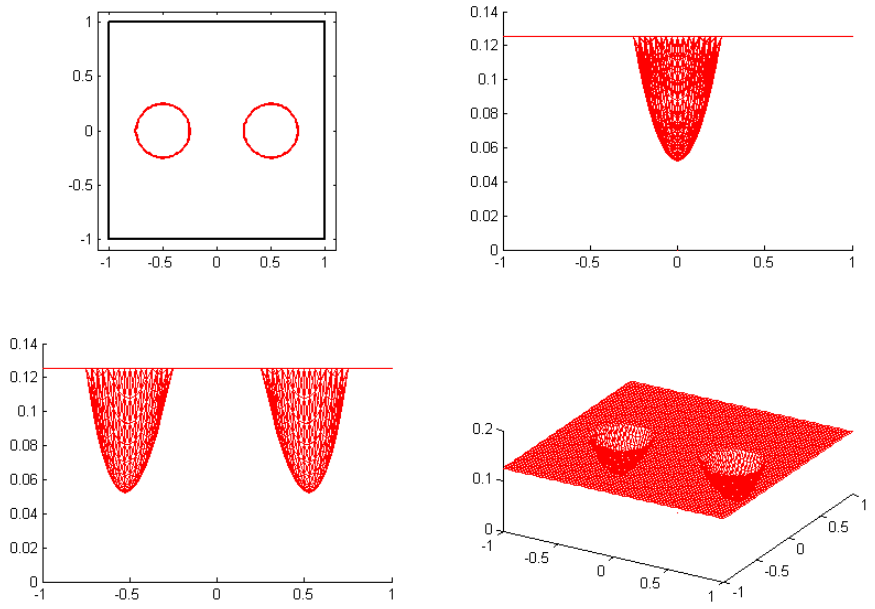


Figure 4-7: *Numerical AMsFEM solution with $H = 1/32$ for multiple inclusion case with $\mathcal{A}_0 = 10^5$ and $\mathcal{A}_1 = 1$. The figure shows the XY-plane and the interfaces (top left), the YZ-projection (top right), the XZ-projection (bottom left) and a 3D view with Z in the vertical direction (bottom right).*

As we can see from Figure 4-7 the case when $\mathcal{A}_0 \to \infty$ is of most interest because the jump in the gradient of the solution across the interface is severe. Therefore it is more interesting to test the standard FEM against the AMsFEM in this poor situation. Such a large jump in gradient does not occur in the case when $\mathcal{A}_1 \to \infty$. The numerical

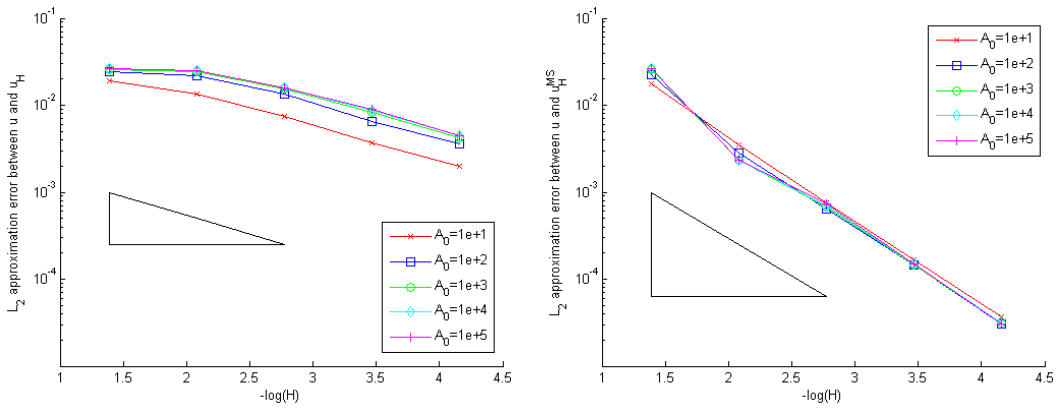
results are shown in Table 4.5 for the standard FEM and Table 4.6 for the Adaptive MsFEM.

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	1.9261e-02	2.4620e-02	2.5911e-02	2.6052e-02	2.6065e-02
1/8	1.3551e-02	2.1990e-02	2.4396e-02	2.4703e-02	2.4734e-02
1/16	7.3986e-03	1.3396e-02	1.5479e-02	1.5769e-02	1.5799e-02
1/32	3.7045e-03	6.5586e-03	8.1858e-03	8.7871e-03	8.8778e-03
1/64	1.9631e-03	3.5752e-03	4.2716e-03	4.4275e-03	4.4458e-03
Rate	0.84601	0.73129	0.67768	0.66049	0.65814

Table 4.5: *L_2 norm of the error using the standard finite element method where $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ for the multiple inclusion experiment.*

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	1.7612e-02	2.2225e-02	2.5670e-02	2.6053e-02	2.6078e-02
1/8	3.4863e-03	2.8106e-03	2.3401e-03	2.3192e-03	2.3628e-03
1/16	7.4512e-04	6.3713e-04	6.7710e-04	6.8998e-04	7.2488e-04
1/32	1.6726e-04	1.4560e-04	1.4775e-04	1.5127e-04	1.5107e-04
1/64	3.7182e-05	3.0982e-05	3.1331e-05	3.1290e-05	3.0555e-05
Rate	2.2157	2.3244	2.3342	2.3341	2.3442

Table 4.6: *L_2 norm of the error for the multiple inclusion experiment using AMsFEM with an immersed FEM subgrid solve, $h = 1/4096$ and where $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$.*



(a) Standard FEM $\mathcal{A}_0 \to \infty$ case.

(b) AMsFEM $\mathcal{A}_0 \to \infty$ case.

Figure 4-8: *Log plots of the standard FEM errors in Table 4.5 against $-log(H)$ (4-8(a)) as well as the adaptive MsFEM errors in Table 4.6 (4-8(b)).*

The results show that the enhanced ALG-MsFEM again outperforms the standard FEM with an optimal rate of convergence that is independent of the contrast parameter. The two inclusions are still comparatively far apart. We will consider later in Section 4.6.4 what happens when inclusions get close together.

4.6.3 Non smooth interfaces

In Assumption 2.20 we assumed that the interfaces have a smooth boundary in order to make use of the regularity result in Theorem 2.22. No such assumption is required to implement the enhanced ALG-MsFEM algorithm. We demonstrate robustness even when there is a singularity present in the following experiment where the inclusion takes the shape of a lens (as seen in Figure 4-9). For this experiment let $r_0 = \pi/6.28$ and $\theta = \pi/4$, then define

$$r = r_0\sqrt{2/(1-\cos(\theta))} \qquad \text{and} \qquad c_y = r_0\sqrt{(1+\cos(\theta))/(1-\cos(\theta))}\ .$$

We then define the lens as the intersection of two circles with radius r and centres $c_1 = (0, -c_y)$, $c_2 = (0, c_y)$ given by

$$\Omega_1 = \{\|x - c_1\| < r\} \cap \{\|x - c_2\| < r\} \qquad \Omega_0 = \Omega \backslash \Omega_1\ .$$

We again use the same load function f boundary conditions g as in the previous examples. The exact solution is unknown so a very fine mesh for the AMsFEM was used, for this $H = 1/128$ and $h = 1/4096$.

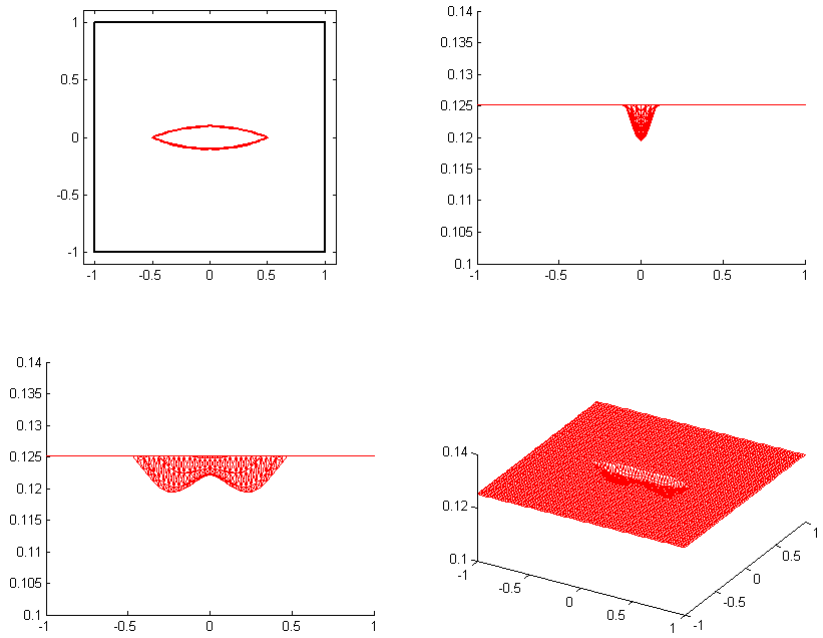


Figure 4-9: *Numerical AMsFEM solution with $H = 1/32$ for the single lens experiment when $\mathcal{A}_0 = 10^5$ and $\mathcal{A}_1 = 1$. The figure shows XY-, YZ-, XZ- and XYZ- projections as in Figure 4-7.*

Note that we have specifically chosen the end points of the lens shape to occur at $(\pm r_0, 0)$ and therefore they will not line up with any uniform mesh with rational H, i.e. the point of singularity is never resolved.

H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	7.7928e-02	8.4253e-02	8.6947e-02	8.7375e-02	8.8468e-02
1/8	2.0201e-02	2.5063e-02	2.6764e-02	2.7098e-02	2.8727e-02
1/16	5.6211e-03	9.5640e-03	1.0771e-02	1.1179e-02	1.3836e-02
1/32	1.8643e-03	4.5901e-03	5.5055e-03	6.1037e-03	8.8843e-03
1/64	7.8831e-04	2.3963e-03	2.9468e-03	3.6807e-03	5.3341e-03
Rate	1.6692	1.2721	1.2047	1.1289	0.97968

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	7.7003e-03	1.3885e-03	1.2077e-03	1.2100e-03	1.2104e-03
1/8	2.0632e-03	1.1681e-03	1.2046e-03	1.2098e-03	1.2102e-03
1/16	7.7087e-04	8.8326e-04	9.4194e-04	9.9737e-04	1.0471e-03
1/32	3.7086e-04	5.8906e-04	6.7075e-04	7.3346e-04	7.6288e-04
1/64	1.9463e-04	3.3039e-04	3.8286e-04	4.0231e-04	4.0647e-04
Rate	1.3088	0.51302	0.41593	0.38993	0.38143

Table 4.7: *L_2 norm of the error using the standard finite element method where $\mathcal{A}_0 = 1$ and $\mathcal{A}_1 \to \infty$ (top) and $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ (bottom)*

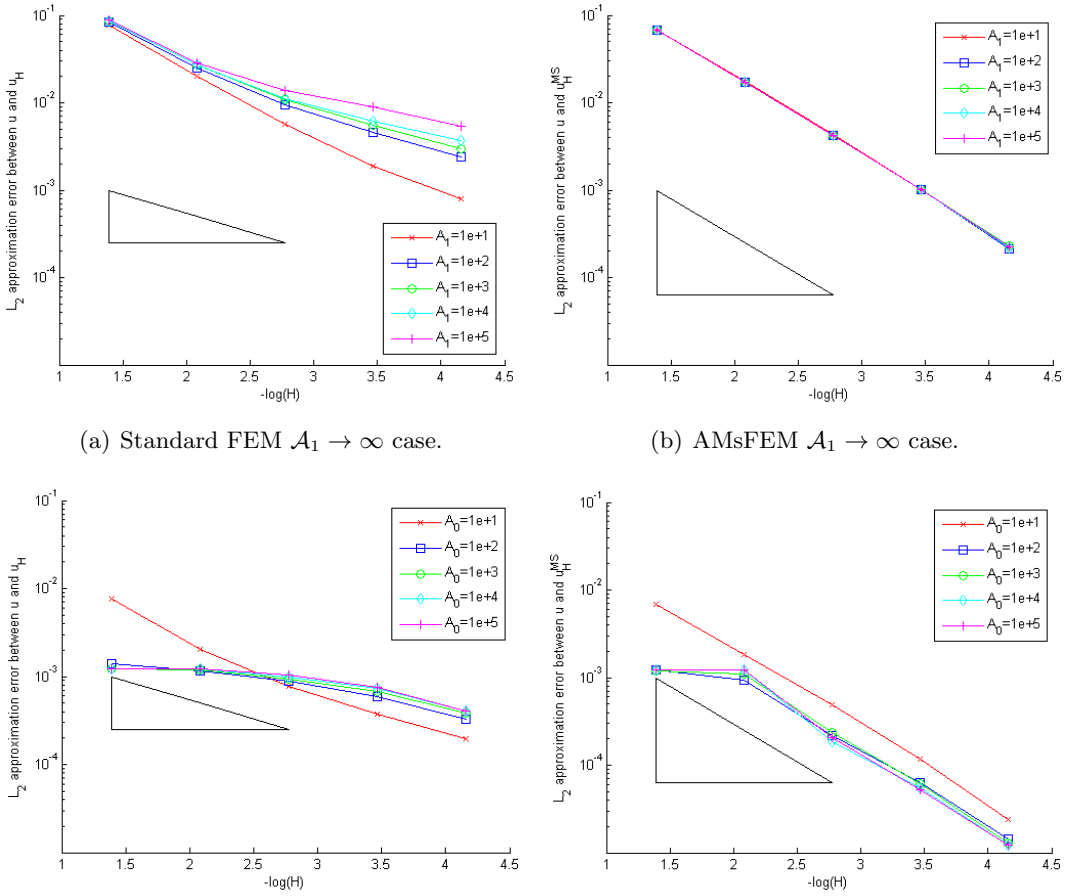
H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	6.8164e-02	6.8200e-02	6.8252e-02	6.8281e-02	6.8220e-02
1/8	1.7416e-02	1.7231e-02	1.7343e-02	1.7317e-02	1.7305e-02
1/16	4.3061e-03	4.2518e-03	4.2503e-03	4.2376e-03	4.2189e-03
1/32	1.0270e-03	1.0108e-03	1.0092e-03	1.0111e-03	1.0188e-03
1/64	2.1133e-04	2.1298e-04	2.2750e-04	2.2041e-04	2.2224e-04
Rate	2.0751	2.0737	2.0561	2.0649	2.0610

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	6.7947e-03	1.2386e-03	1.1948e-03	1.2102e-03	1.2098e-03
1/8	1.8208e-03	9.3799e-04	1.0915e-03	1.2099e-03	1.2107e-03
1/16	4.9382e-04	2.2109e-04	2.3627e-04	1.8742e-04	2.0840e-04
1/32	1.1688e-04	6.2451e-05	6.2075e-05	5.4692e-05	5.2183e-05
1/64	2.4316e-05	1.4255e-05	1.2914e-05	1.2336e-05	1.2220e-05
Rate	2.0214	1.6791	1.7199	1.7700	1.7795

Table 4.8: *L_2 norm of the error using AMsFEM with an immersed FEM subgrid solve where $\mathcal{A}_0 = 1$ and $\mathcal{A}_1 \to \infty$ (top) and $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ (bottom) with $M = 32$.*

The results in Tables 4.7 and 4.8 show that the standard FEM performs as expected in the first instance when $\mathcal{A}_1 \to \infty$ while it performs very poorly when $\mathcal{A}_0 \to \infty$. In contrast the enhanced ALG-MsFEM performs well in both cases with only a slight drop in convergence rate when $\mathcal{A}_0 \to \infty$. An important observation is that the standard FEM is starting to exhibit contrast dependent behaviour in the first case where the convergence rate is rapidly falling away with increasing $\mathcal{A}_1$ while the enhanced

ALG-MsFEM remains unaffected. This experiment starts to show the strength of the enhanced ALG-MsFEM when applied to problems that contain a singularity.



(a) Standard FEM $\mathcal{A}_1 \to \infty$ case.

(b) AMsFEM $\mathcal{A}_1 \to \infty$ case.

(c) Standard FEM $\mathcal{A}_0 \to \infty$ case.

(d) AMsFEM $\mathcal{A}_0 \to \infty$ case.

Figure 4-10: *Log plots of the standard FEM errors in Table 4.7 against $-log(H)$ (graphs 4-10(a) and 4-10(c)) as well as the adaptive MsFEM errors in Table 4.8 (graphs 4-10(b) and 4-10(d)).*

Following on from the single lens experiment we consider a double lens experiment with two lens next to each other. We shift the lenses slightly to the right so that the cross point does not fall on the coarse mesh. For this experiment we only consider the $\mathcal{A}_0 \to \infty$ case since this is the case with a significant jump in gradient and thus gives a harder test to compare the FEM to AMsFEM. We use the same load function f and boundary conditions g but redefine the coefficient. Let $r_0 = \pi/12.56$, $\theta = \pi/4$ and define

$$r = r_0\sqrt{2/(1-\cos(\theta))}, \qquad c_y = r_0\sqrt{(1+\cos(\theta))/(1-\cos(\theta))} \qquad \text{and} \qquad \epsilon = 1/128\ .$$

Then for $c_1 = (-r_0 + \epsilon, -c_y)$, $c_2 = (-r_0 + \epsilon, c_y)$, $c_3 = (r_0 + \epsilon, -c_y)$ and $c_4 = (r_0 + \epsilon, c_y)$ let

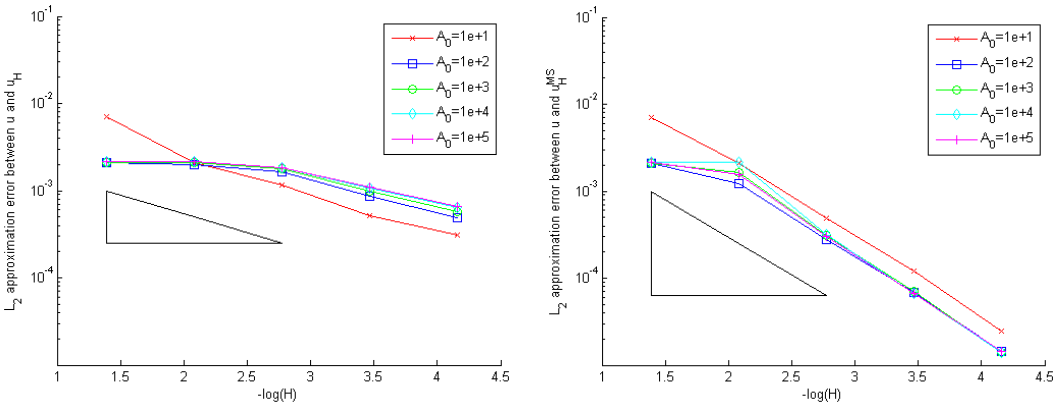
$$\Omega_1 = (\{\|x - c_1\| < r\} \cap \{\|x - c_2\| < r\}) \cup (\{\|x - c_3\| < r\} \cap \{\|x - c_4\| < r\}) \ ,$$
$$\Omega_0 = \Omega \backslash (\Omega_1 \cap \Omega_2) \ .$$

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	7.0502e-03	2.0726e-03	2.1193e-03	2.1352e-03	2.1369e-03
1/8	2.0904e-03	2.0135e-03	2.1235e-03	2.1356e-03	2.1368e-03
1/16	1.1454e-03	1.6439e-03	1.7980e-03	1.8320e-03	1.8360e-03
1/32	5.1831e-04	8.6355e-04	9.9545e-04	1.0763e-03	1.1011e-03
1/64	3.0998e-04	4.9551e-04	5.8234e-04	6.4139e-04	6.5887e-04
Rate	1.1027	0.53503	0.48203	0.44588	0.43514

Table 4.9: *L_2 norm of the error using the standard finite element method where $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$.*

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	6.9574e-03	2.1056e-03	2.1223e-03	2.1349e-03	2.1364e-03
1/8	2.0970e-03	1.2313e-03	1.6277e-03	2.1290e-03	1.5417e-03
1/16	4.9075e-04	2.7460e-04	3.0800e-04	3.2111e-04	3.0485e-04
1/32	1.2018e-04	6.9102e-05	6.9817e-05	6.6707e-05	6.5948e-05
1/64	2.4726e-05	1.4326e-05	1.3826e-05	1.3846e-05	1.4193e-05
Rate	2.0398	1.8554	1.9067	1.9533	1.9015

Table 4.10: *L_2 norm of the error using AMsFEM with an immersed FEM subgrid solve where $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ with $M = 32$.*



(a) Standard FEM $\mathcal{A}_0 \to \infty$ case.

(b) AMsFEM $\mathcal{A}_0 \to \infty$ case.

Figure 4-11: *Log plots of the standard FEM errors in Table 4.9 against $-log(H)$ (4-11(a)) as well as the adaptive MsFEM errors in Table 4.10 (4-11(b)).*

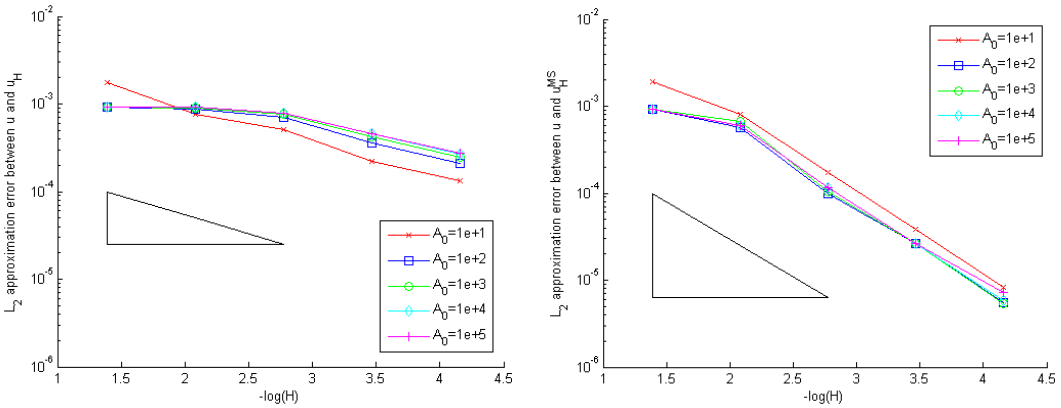
We repeat the double lens experiment again but this time with homogeneous boundary data ($g = 0$) and load function $f = 1$. This is to show that it is not the specific choice of boundary conditions or load functions that is giving the superior convergence rates. Again we consider only the case when $\mathcal{A}_0 \to \infty$ and the results are displayed in Tables 4.11 and 4.12 as well as graphically in Figure 4-12.

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	1.7711e-03	9.2178e-04	9.1320e-04	9.1351e-04	9.1355e-04
1/8	7.6700e-04	8.7650e-04	9.0953e-04	9.1310e-04	9.1346e-04
1/16	5.0995e-04	7.0501e-04	7.6492e-04	7.7835e-04	7.7996e-04
1/32	2.2325e-04	3.6206e-04	4.1780e-04	4.5277e-04	4.6335e-04
1/64	1.3200e-04	2.0798e-04	2.4481e-04	2.6938e-04	2.7617e-04
Rate	0.92726	0.55715	0.49208	0.45356	0.44311

Table 4.11: *L_2 norm of the error using the standard finite element method where $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$.*

H	$\mathcal{A}_0$ = 1e+1	$\mathcal{A}_0$ = 1e+2	$\mathcal{A}_0$ = 1e+3	$\mathcal{A}_0$ = 1e+4	$\mathcal{A}_0$ = 1e+5
1/4	1.9096e-03	9.2448e-04	9.1297e-04	9.1330e-04	9.1334e-04
1/8	7.9571e-04	5.6258e-04	6.7492e-04	5.9817e-04	6.0590e-04
1/16	1.7430e-04	9.7538e-05	1.0542e-04	1.1651e-04	1.1635e-04
1/32	3.8673e-05	2.6721e-05	2.6248e-05	2.6028e-05	2.6291e-05
1/64	8.1844e-06	5.4943e-06	5.4405e-06	5.8006e-06	7.2820e-06
Rate	2.0095	1.9185	1.9466	1.9120	1.8468

Table 4.12: *L_2 norm of the error using AMsFEM with an immersed FEM subgrid solve where $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ with $M = 32$.*



(a) Standard FEM $\mathcal{A}_0 \to \infty$ case.

(b) AMsFEM $\mathcal{A}_0 \to \infty$ case.

Figure 4-12: *Log plots of the standard FEM errors in Table 4.11 against $-log(H)$ (4-12(a)) as well as the adaptive MsFEM errors in Table 4.12 (4-12(b)).*

In both of the previous experiments we see that the standard FEM is struggling to

converge with the cross point and the end points of the double lens present but the enhanced ALG-MsFEM has no difficulty. It shows that the method is very versatile for a wide range of permeability fields even if the coarse mesh $\mathcal{T}_H(\Omega)$ does not resolve the interface and particularly if it does not resolve singularity points.

4.6.4 Boundary layer interfaces

In this section we explore the effectiveness of the enhanced ALG-MsFEM when the inclusions get close to the boundary and close to each other. This shows that the ALG-MsFEM is not subject to the Assumption 2.20 which we needed for the analysis in Chapter 2. Here we consider an oval inclusion whos top and bottom edges approach the boundary $\partial\Omega$ (see Figure 4-13). This example also includes Neumann boundary conditions. Let $r_0 = 1 - \epsilon$ and $\epsilon = 1/32$, then $r = (2x)^4 + y^4 - r_0^4$ giving

$$\Omega_1 = \{x \in \Omega \,|\, r(x) < 0\} \qquad \Omega_0 = \Omega \backslash \Omega_1 \ .$$

We also take Dirichlet boundary conditions on $\Gamma_D = \{(x, y) \in \partial\Omega \,|\, y = -1, 1\}$ and no-flow Neumann conditions on $\Gamma_N = \partial\Omega \backslash \Gamma_D$. We use the data

$$f = 0 \ , \qquad g|_{y=-1} = 0 \ , \qquad g|_{y=1} = 1 \ .$$

The exact solution is unknown so to obtain a reference solution a very fine mesh for the AMsFEM was used, for this $H = 1/128$ and $h = 1/4096$.

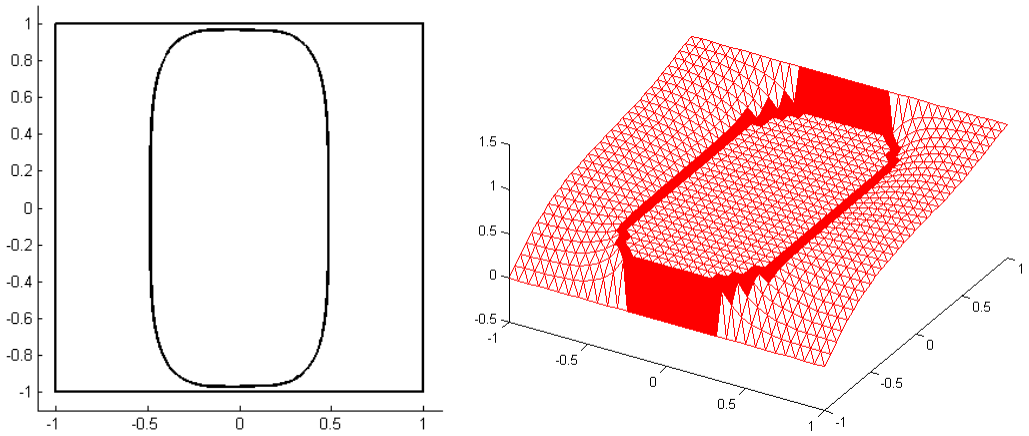


Figure 4-13: *The inclusions Ω_0 and Ω_1 for the epsilon boundary layer experiment with $\epsilon = 1/32$ (left) and an example solution from the adaptive MsFEM for the case when $\mathcal{A}_0 = 10^5$ and $H = \frac{1}{16}$ (right). The adaptive MsFEM uses a subgrid on cut elements with $h = \frac{1}{512}$.*

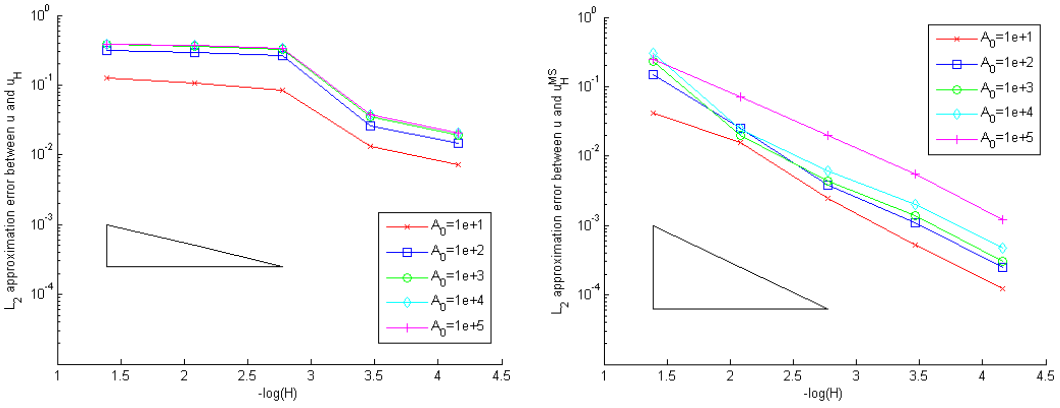
H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	1.2504E-01	3.1627E-01	3.7827E-01	3.8589E-01	3.8655E-01
1/8	1.0712E-01	2.9696E-01	3.5846E-01	3.6602E-01	3.6667E-01
1/16	8.5114E-02	2.6780E-01	3.2785E-01	3.3524E-01	3.3585E-01
1/32	1.3348E-02	2.5922E-02	3.4751E-02	3.7658E-02	3.7702E-02
1/64	7.3041E-03	1.4669E-02	1.9193E-02	2.0472E-02	2.0432E-02
Rate	1.1200	1.2379	1.1968	1.1754	1.1765

Table 4.13: *L_2 norm of the error using the standard finite element method where $\mathcal{A}_1 \to \infty$ and $\mathcal{A}_0 = 1$.*

H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	4.1972E-02	1.4734E-01	2.3291E-01	3.0717E-01	2.4719E-01
1/8	1.5822E-02	2.5200E-02	1.9493E-02	2.3984E-02	7.1733E-02
1/16	2.4185E-03	3.8372E-03	4.3718E-03	6.1675E-03	1.9729E-02
1/32	5.1496E-04	1.1014E-03	1.3663E-03	2.0203E-03	5.4806E-03
1/64	1.2160E-04	2.4638E-04	3.0904E-04	4.7413E-04	1.2130E-03
Rate	2.1804	2.2964	2.2950	2.2248	1.9052

Table 4.14: *L_2 norm of the error using AMsFEM with an immersed FEM subgrid solve where $\mathcal{A}_1 \to \infty$ and $\mathcal{A}_0 = 1$ with $M = 32$.*

The numerical results are shown in Tables 4.13 and 4.14 but the convergence rates calculated through linear regression do not show the true impact of the enhanced ALG-MsFEM over the standard FEM. A better representation of the results is given via the graphs in Figure 4-14.



(a) Standard FEM $\mathcal{A}_1 \to \infty$ case.

(b) AMsFEM $\mathcal{A}_1 \to \infty$ case.

Figure 4-14: *Log plots of the standard FEM errors in Table 4.13 against $-log(H)$ (4-14(a)) as well as the adaptive MsFEM errors in Table 4.14 (4-14(b)).*

The graph on the left shows the results for the standard FEM. We can clearly see that there is a boundary layer effect present where the standard FEM converges very

slowly whilst the mesh size H is larger than $\epsilon = 1/32$ but then speeds up when H gets smaller. No such boundary layer problem exists when the enhanced ALG-MsFEM is used showing a significant advantage when boundary layers are present. We also note that introducing mixed boundary conditions does not pose a restriction to the enhanced ALG-MsFEM or affect its convergence rate.

To explore the effect of boundary layers further we consider several inclusions that are close together and close to the boundary. The inclusions take the form of four ovals that are a distance ϵ from the boundary and 2ϵ from each other in an arrangement that, using the same f and g as the previous experiment, gives a solution with a series of steps (see Figure 4-15).

For this experiment let $r_x = \frac{1}{2} - \epsilon$, $r_y = \frac{1}{4} - \epsilon$ and $\epsilon = 1/32$, then

$$r = \min_{p=-3,-1,1,3} \left(\frac{x}{r_x} \right)^4 + \left(\frac{y - \frac{p}{4}}{r_y} \right)^4 - 1 \ .$$

Thus

$$\Omega_1 = \{x \in \Omega \,|\, r < 0\} \qquad \Omega_0 = \Omega \backslash \Omega_1 \ .$$

We utilise the same f, g, Γ_D and Γ_N as in the single epsilon boundary layer experiment previously. The exact solution is unknown so to obtain a reference solution a very fine mesh for the AMsFEM was used, for this $H = 1/128$ and $h = 1/4096$.

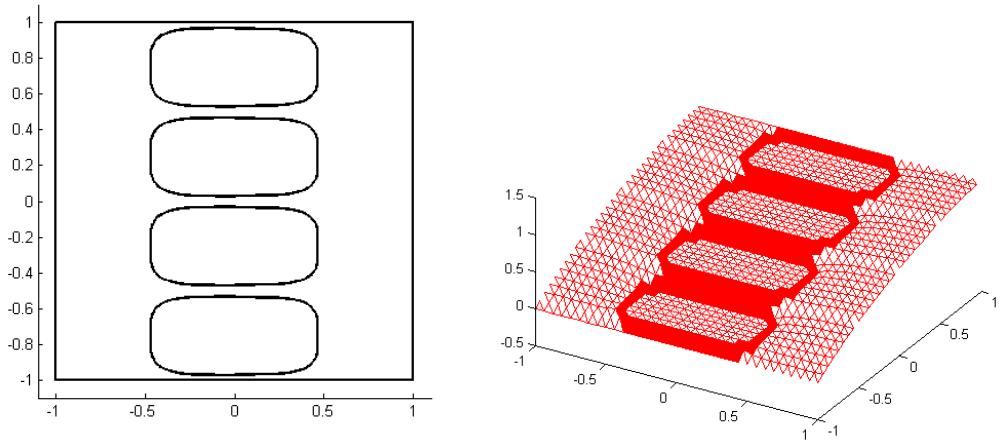


Figure 4-15: *The inclusions Ω_0 and Ω_1 for the multiple epsilon boundary layer experiment with $\epsilon = 1/32$ (left) and an example solution from the adaptive MsFEM for the case when $\mathcal{A}_0 = 10^5$ and $H = \frac{1}{16}$ (right). The adaptive MsFEM uses a subgrid on cut elements with $h = \frac{1}{512}$.*

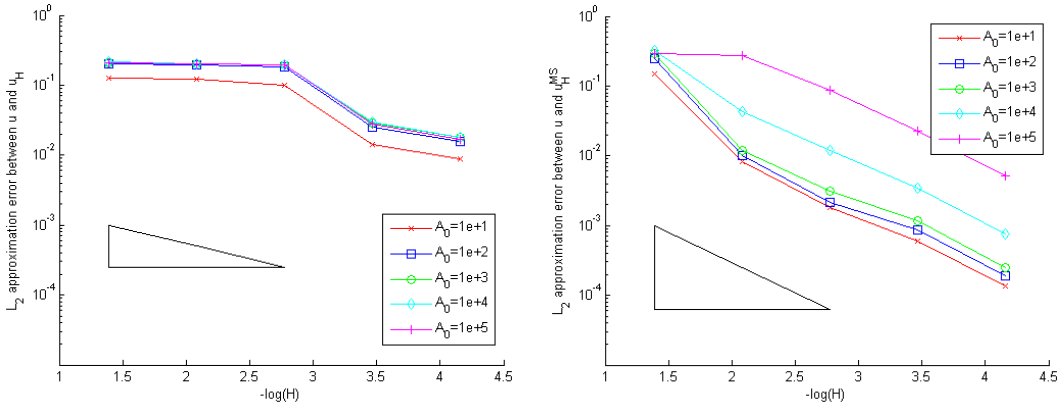
H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	1.2574E-01	2.0134E-01	2.1209E-01	2.1304E-01	2.1161E-01
1/8	1.2370E-01	1.9411E-01	2.0404E-01	2.0491E-01	2.0344E-01
1/16	1.0096E-01	1.8234E-01	1.9592E-01	1.9743E-01	1.9601E-01
1/32	1.4049E-02	2.4855E-02	2.8688E-02	2.9184E-02	2.7869E-02
1/64	8.6962E-03	1.5599E-02	1.7904E-02	1.8187E-02	1.6734E-02
Rate	1.0846	1.0345	0.9963	0.9912	1.0189

Table 4.15: *L_2 norm of the error using the standard finite element method where $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$.*

H	$\mathcal{A}_1$ = 1e+1	$\mathcal{A}_1$ = 1e+2	$\mathcal{A}_1$ = 1e+3	$\mathcal{A}_1$ = 1e+4	$\mathcal{A}_1$ = 1e+5
1/4	4.1972E-02	1.4734E-01	2.3291E-01	3.0717E-01	2.4719E-01
1/8	1.5822E-02	2.5200E-02	1.9493E-02	2.3984E-02	7.1733E-02
1/16	2.4185E-03	3.8372E-03	4.3718E-03	6.1675E-03	1.9729E-02
1/32	5.1496E-04	1.1014E-03	1.3663E-03	2.0203E-03	5.4806E-03
1/64	1.2160E-04	2.4638E-04	3.0904E-04	4.7413E-04	1.2130E-03
Rate	2.4007	2.4293	2.3737	2.1139	1.5228

Table 4.16: *L_2 norm of the error using AMsFEM with an immersed FEM subgrid solve where $\mathcal{A}_0 \to \infty$ and $\mathcal{A}_1 = 1$ with $M = 32$.*

The results again show a boundary layer effect while the coarse mesh diameter H is greater than $\epsilon = 1/32$. The disadvantage now is that while the rate of the enhanced ALG-MsFEM is still good the size of the error is starting to depend on the size of the contrast in the coefficient. This is because the gradient between the inclusions is extreme as the contrast increases. Further investigation in this extreme circumstance is required but the enhanced ALG-MsFEM is still superior to the standard FEM.



(a) Standard FEM $\mathcal{A}_1 \to \infty$ case.

(b) AMsFEM $\mathcal{A}_1 \to \infty$ case.

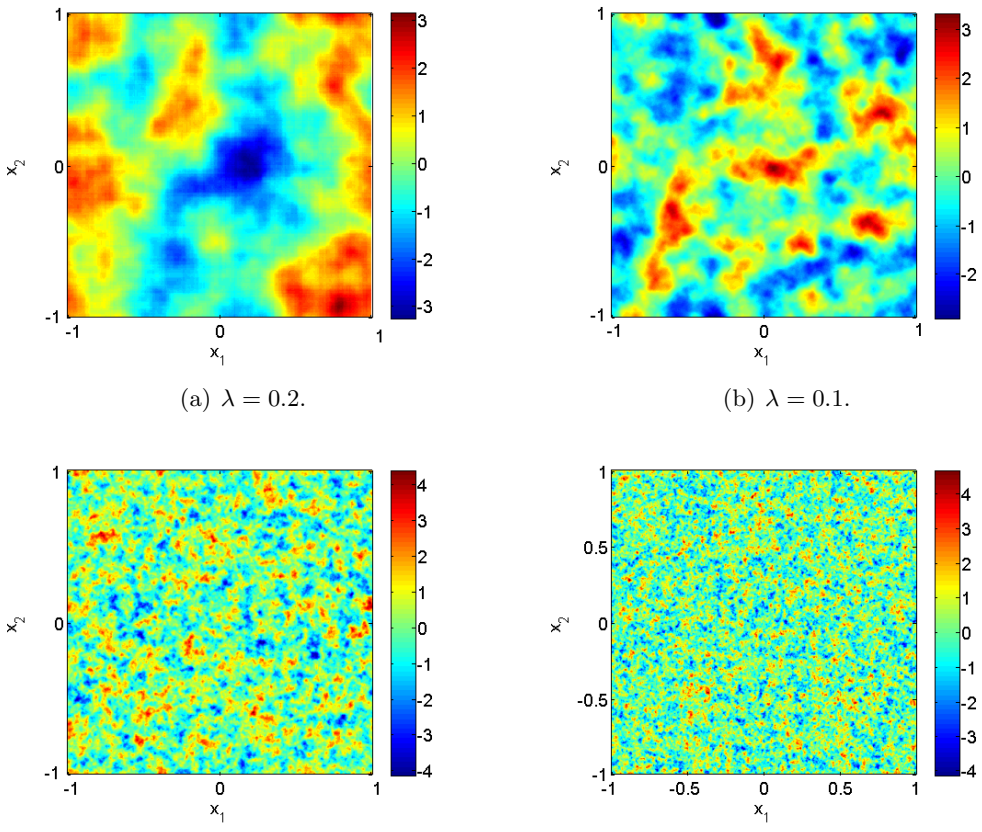
Figure 4-16: *Log plots of the standard FEM errors in Table 4.15 against $-log(H)$ (4-16(a)) as well as the adaptive MsFEM errors in Table 4.16 (4-16(b)).*

4.6.5 Random field problems

For the last set of experiments we show the full generality of the enhanced ALG-MsFEM by considering a coefficient $\mathcal{A}$ that is not defined as a set of inclusions. Instead the coefficient is given by a matrix of values representing a log normal permeability field for a rock structure. We find the permeability field $\mathcal{A}(x)$ by first defining $\widetilde{Z}(x,w)$ as a Gaussian random field with mean $\mu = 0$ and standard deviation $\sigma = 1$ for $x \in \Omega$. The random field $\widetilde{Z}(x,w)$ satisfies the covariance function

$$\mathbb{E}\left[\widetilde{Z}(x,w), \widetilde{Z}(y,w)\right] = \sigma^2 \exp\left(-\left\|x-y\right\|_2 / \lambda\right) \ ,$$

where λ is the length scale. We can then define random fields with different standard deviations by setting $Z(x,w) = \sigma\widetilde{Z}(x,w)$. Finally we obtain the permeability field by setting $\mathcal{A}(x,w) = \exp\left(Z(x,w)\right)$.



(a) $\lambda = 0.2$. (b) $\lambda = 0.1$.

(c) $\lambda = 0.02$. (d) $\lambda = 0.01$.

Figure 4-17: *The random fields used for this experiment. The images show the Guassian random field $\widetilde{Z}(x,w)$ with zero mean and standard deviation $\sigma = 1$ for decreasing values of the length scale λ.*

The random fields $\widetilde{Z}(x, w)$ used in the following experiments are shown in Figure 4-17. They constitute single events and the following experiments are designed simply to show that the enhanced ALG-MsFEM can be used effectively for problems with very heterogeneous coefficients. A much more extensive statistical study needs to be performed to fully examine the performance of ALG-MsFEM for these types of problems but the following experiments give promising initial results.

In our first experiment we consider the effect of decreasing the length scale λ whilst maintaining a constant variance $\sigma^2 = 1$. This gives a moderate contrast in the coefficient of the order of 10^3 but still poses an effective test. This first experiment uses a load function $f = 1$ and zero Dirichlet boundary conditions $g = 0$ on $\Gamma_D := \partial\Omega$ (i.e. recharging boundary conditions). No exact solution exists so we compare the approximate solutions to a reference solution computed using the standard FEM on a fine grid with $h_{\text{fine}} = 1/128$ for $\lambda = 0.2, 0.1$, $h_{\text{fine}} = 1/512$ for $\lambda = 0.02$ and $h_{\text{fine}} = 1/1024$ for $\lambda = 0.01$. The results are then displayed in Tables 4.17 and 4.18.

H	λ = 2e-1	λ = 1e-1	λ = 2e-2	λ = 1e-2
1/4	5.5911E-03	9.1245E-03	1.3894E-02	1.3699E-02
1/8	2.0674E-03	3.9063E-03	1.1498E-02	1.2627E-02
1/16	6.4957E-04	1.4951E-03	8.5419E-03	1.0956E-02
1/32	1.8243E-04	4.8258E-04	4.7580E-03	8.2113E-03
1/64	4.3003E-05	1.3718E-04	2.0124E-03	4.6453E-03
Rate	1.7548	1.5128	0.6848	0.3741

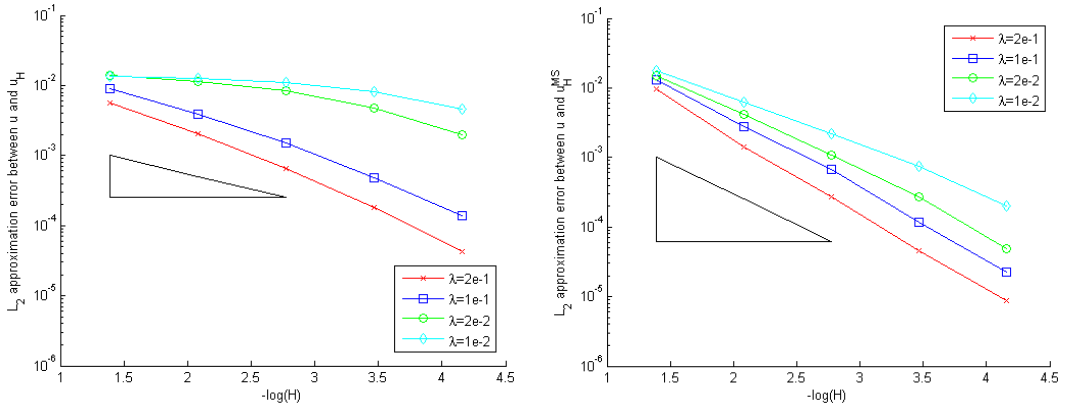
Table 4.17: *L_2 norm of the error using the standard finite element method where the standard deviation $\sigma = 1$.*

H	λ = 2e-1	λ = 1e-1	λ = 2e-2	λ = 1e-2
1/4	9.6130E-03	1.3159E-02	1.5170E-02	1.7693E-02
1/8	1.4371E-03	2.7423E-03	4.0918E-03	6.2758E-03
1/16	2.6858E-04	6.7361E-04	1.0821E-03	2.2158E-03
1/32	4.5650E-05	1.1710E-04	2.7516E-04	7.3929E-04
1/64	8.8451E-06	2.2147E-05	4.9223E-05	1.9915E-04
Rate	2.5148	2.2979	2.0430	1.6032

Table 4.18: *L_2 norm of the error using AMsFEM with an standard FEM subgrid solve where the standard deviation $\sigma = 1$ with $h = H/8$.*

The results are shown graphically in Figure 4-18 where we see that the convergence rate of the standard FEM is heavily dependent on the length scale. This is expected as the change in $\mathcal{A}$ is on a much smaller scale than the coarse mesh diameter H. The interesting result from this experiment is that the enhanced ALG-MsFEM, while mildly

dependent on the length scale, converges at an optimal rate but with a subgrid element size $h = H/8$ which does not always resolve the length scale. Generally the standard method requires a mesh diameter of size about $\lambda/10$ in order to resolve the length scale. In the enhanced ALG-MsFEM however even the subgrid size h still does not reach this level. Particularly in the case of $\lambda = 0.01$, $h = H/8$ is much larger than $\lambda/10$. Even at the finest level then $h = \lambda/5$.



(a) Standard FEM $\lambda \to 0$ case.

(b) AMsFEM $\lambda \to 0$ case.

Figure 4-18: *Log plots of the standard FEM errors in Table 4.17 against $-log(H)$ (4-18(a)) as well as the adaptive MsFEM errors in Table 4.18 (4-18(b)).*

Our next experiment seeks to consider the effect of increasing the standard deviation σ which effectively increases the contrast in the coefficient $\mathcal{A}$. To make the experiment as extreme as possible we consider a small length scale $\lambda = 0.01$ with the random field $\widetilde{Z}(x, w)$ as shown in Figure 4-17(d). We then produce a new field with different standard deviations by letting $Z(x, w) = \sigma\widetilde{Z}(x, w)$ which gives us the same field structure but with greater contrast. Varying σ gives the following contrast values:

σ	0.5	1.0	1.5	2.0	2.5
$\mathcal{A}_{\max}/\mathcal{A}_{\min}$	8.0746E+01	6.5200E+03	5.2646E+05	4.2510E+07	3.4325E+09

For the experiment with varying σ we again use the same load function $f = 1$ and boundary conditions $g = 0$ as the λ experiment. No exact solution exists so we obtain a reference solution using the standard FEM on a fine grid with $h_{\text{fine}} = 1/1024$. The numerical results in Table 4.19 show that again the standard FEM converges very poorly with an error that is growing with increasing contrast. However, in this extreme test the enhanced ALG-MsFEM is converging slower than in previous experiments. What can be seen graphically in Figure 4-19 is that the rate of convergence is accelerating to $O(H^2)$, taking longer to do so as the contrast increases. We note again that this ex-

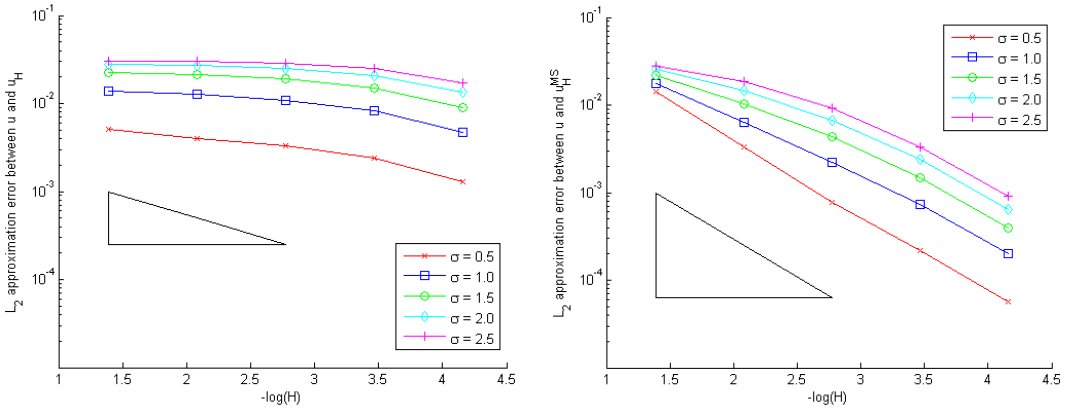
periment also uses a subgrid $h = H/8$ and therefore even the fine mesh in each element does not resolve the length scale sufficiently (regarded as $\lambda/10$) thus again showing the the enhanced ALG-MsFEM is a very powerful tool for difficult problems, for example problems that involve highly varying random fields.

H	σ = 0.5	σ = 1.0	σ = 1.5	σ = 2.0	σ = 2.5
1/4	5.0650E-03	1.3699E-02	2.2454E-02	2.8078E-02	3.0475E-02
1/8	3.9891E-03	1.2627E-02	2.1435E-02	2.7248E-02	2.9947E-02
1/16	3.2794E-03	1.0956E-02	1.9103E-02	2.5054E-02	2.8408E-02
1/32	2.3710E-03	8.2113E-03	1.5004E-02	2.0834E-02	2.4973E-02
1/64	1.2975E-03	4.6453E-03	8.9629E-03	1.3360E-02	1.7309E-02
Rate	0.4680	0.3741	0.3164	0.2530	0.1894

Table 4.19: *L_2 norm of the error using the standard finite element method where the length scale $\lambda = 0.01$.*

H	σ = 0.5	σ = 1.0	σ = 1.5	σ = 2.0	σ = 2.5
1/4	1.4128E-02	1.7693E-02	2.1917E-02	2.5506E-02	2.7919E-02
1/8	3.3298E-03	6.2758E-03	1.0403E-02	1.4714E-02	1.8623E-02
1/16	7.8395E-04	2.2158E-03	4.2922E-03	6.6902E-03	9.2049E-03
1/32	2.1933E-04	7.3929E-04	1.4870E-03	2.3735E-03	3.3060E-03
1/64	5.7356E-05	1.9915E-04	3.9830E-04	6.3380E-04	9.1038E-04
Rate	1.8864	1.7380	1.7149	1.7000	1.6689

Table 4.20: *L_2 norm of the error using AMsFEM with an standard FEM subgrid solve where the length scale $\lambda = 0.01$ with $M = 8$. (Rates calculated by linear regression over the last three entries per column.)*



(a) Standard FEM $\sigma \to \infty$ case.

(b) AMsFEM $\sigma \to \infty$ case.

Figure 4-19: *Log plots of the standard FEM errors in Table 4.19 against $-log(H)$ (4-19(a)) as well as the adaptive MsFEM errors in Table 4.20 (4-19(b)).*

Our last experiment shows that in the previous experiment the increasing error with respect to increasing contrast is actually due to the presence of the load function

f. Instead for this experiment we consider $f = 0$ and take mixed boundary conditions as in Section 4.6.4 for the boundary layer problems. Therefore the coefficient is defined in the previous experiment, we take Dirichlet boundary conditions on $\Gamma_D = \{(x, y) \in \partial\Omega \,|\, y = -1, 1\}$ and no-flow Neumann conditions on $\Gamma_N = \partial\Omega \backslash \Gamma_D$. We use the data

$$f = 0 \,, \qquad g|_{y=-1} = 0 \,, \qquad g|_{y=1} = 1 \,.$$

No exact solution exists so we take the approximate solution to the standard FEM on a fine grid with $h_{\text{fine}} = 1/1024$.

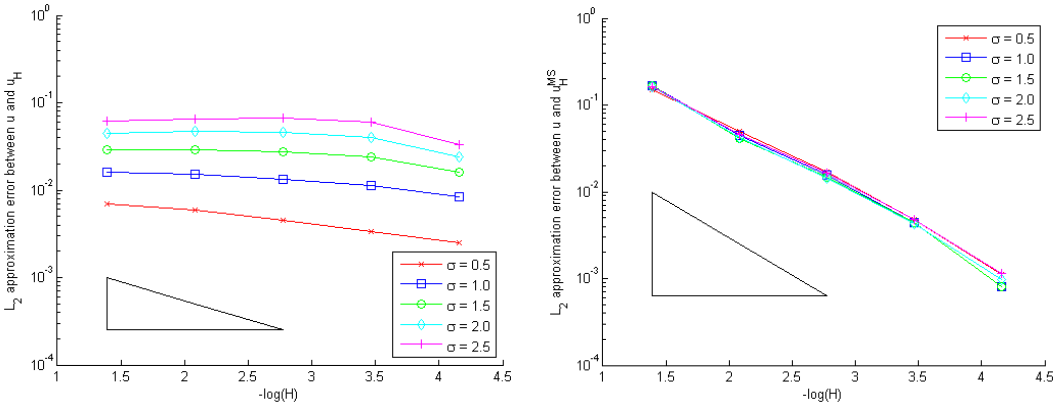
H	σ = 0.5	σ = 1.0	σ = 1.5	σ = 2.0	σ = 2.5
1/4	6.9328E-03	1.6103E-02	2.9291E-02	4.5052E-02	6.1052E-02
1/8	5.8513E-03	1.5004E-02	2.9135E-02	4.6834E-02	6.5094E-02
1/16	4.5167E-03	1.3215E-02	2.7189E-02	4.5438E-02	6.6224E-02
1/32	3.3733E-03	1.1304E-02	2.3849E-02	4.0540E-02	6.0010E-02
1/64	2.4828E-03	8.3413E-03	1.5876E-02	2.4075E-02	3.2935E-02
Rate	0.3758	0.2306	0.2056	0.2016	0.1898

Table 4.21: *L_2 norm of the error using the standard finite element method where the length scale $\lambda = 0.01$.*

H	σ = 0.5	σ = 1.0	σ = 1.5	σ = 2.0	σ = 2.5
1/4	1.5021E-01	1.6490E-01	1.6737E-01	1.6368E-01	1.6033E-01
1/8	5.0160E-02	4.4875E-02	4.1318E-02	4.1895E-02	4.5996E-02
1/16	1.6751E-02	1.5504E-02	1.4708E-02	1.4502E-02	1.6462E-02
1/32	4.7396E-03	4.4478E-03	4.3567E-03	4.2995E-03	4.8243E-03
1/64	1.1010E-03	8.0723E-04	8.0435E-04	9.6450E-04	1.1592E-03
Rate	1.7588	1.8684	1.8647	1.8098	1.7477

Table 4.22: *L_2 norm of the error using AMsFEM with an standard FEM subgrid solve where the length scale $\lambda = 0.01$ with $M = 8$.*

The results for the standard FEM in Table 4.21 again show a very poor convergence rate and the error is growing with the contrast. No such dependence on the contrast exists for the enhanced ALG-MsFEM and the convergence rate is only slightly less than optimal. There is however a drawback here to the enhanced ALG-MsFEM, working on the premise that the L_2 norm finite element error is of the form $C_1 H^{0.3}$ for the standard FEM and $C_2 H^{1.8}$ for the enhanced ALG-MsFEM, we can see that $C_2 > C_1$. This means that the enhanced ALG-MsFEM only beats the standard FEM for sufficiently small H. The important point to remember though is that it still has a much higher rate of convergence and is robust with respect to the contrast parameter, meaning that the enhanced ALG-MsFEM is still a very effective method.



(a) Standard FEM $\sigma \to \infty$ case.

(b) AMsFEM $\sigma \to \infty$ case.

Figure 4-20: *Log plots of the standard FEM errors in Table 4.21 against $-log(H)$ (4-20(a)) as well as the adaptive MsFEM errors in Table 4.22 (4-20(b)).*

4.7 Summary

In this chapter we have shown how to obtain multiscale basis functions iteratively with a method based on the local-global techniques by Durlofsky, Efendiev and Ginting in [36]. This has allowed us to move beyond the interface problem to general high contrast elliptic interface problems with more general boundary conditions. We have described the general algorithm and examined some of the properties of the algorithm. We have shown how the "conforming" (EDG1) and "non-conforming" (EDG2) local-global methods are actually special cases of a more general adaptive multiscale framework and then proposed an enhancement to obtain a good convergence rate ($O(H)$ in the energy norm) given an L_∞ coefficient $\mathcal{A}$ but with a conforming method.

In Section 4.6 we explored many numerical examples to show the improvement of the adaptive multiscale finite element method over the standard FEM. The improvement was shown to be particularly dramatic in the case when the coefficient contained a corner singularity or a boundary layer. We also saw a significant improvement when the Adaptive MsFEM was used for problems with a log normal random field particularly when the problem had no source term and was driven by the boundary conditions.

In the next chapter we will see how the Adaptive MsFEM can be applied to linear elasticity problems for situations arising in structural engineering as opposed to being commonly used in the field of porous media flow.

5

Application to shape optimisation in linear elasticity

So far in this thesis we have examined the drawbacks of the standard finite element method for scalar elliptic problems but also introduced a new type of local-global method and demonstrated superior convergence. Many of the examples introduced so far have only been model examples designed to show the capability of the adaptive method, its ability to deal with interfaces and provide optimal convergence. In Section 4.6.5 we saw how the enhanced ALG-MsFEM could also be applied to more general heterogeneous problems. We note however that all of these examples were for a scalar elliptic equation.

The assumptions in Chapter 2 were imposed to make the analysis tractable and we have shown experimentally how the enhanced ALG-MsFEM does not require such restrictions. Much of the work on local-global methods, in fact most multiscale work, has been applied to problems such as porous media flow. The adaptive method, in the general form presented in Chapter 4, can also be applied to problems in linear elasticity. In this chapter, we examine how the adaptive method can aid in solving problems in structural optimisation.

Structural optimisation is used by engineers to find the best structure to minimise a cost function. For example they may want to find the stiffest structure to hold a load off the ground. The solution to this would be to place the object on a solid block. However, in a world of limited resources and limited costs it is also important to find the structure that uses least material and this is where the difficulty of structural optimisation comes in. The idea is to start with an initial structure and solve a linear elasticity problem (see Section 5.2), using the resulting solution, the interface is then moved to obtain a new structure. The process is repeated until a sufficiently optimal solution is obtained.

The chapter will start by generalizing the problem definition and associated notation used in Chapter 2. Instead we will introduce an abstract variational problem in terms

of a bounded and coercive bilinear form that depends on an L_∞ coefficient. We then explore how the linear elasticity problem is defined. We state the structural optimisation process where the linear elasticity problem is solved at each step with a different structure. Using this newly obtained approximation of the displacement under loading conditions we define a shape sensitivity indicator for moving the boundary of the current structure to obtain a new structure (the shape sensitivity value at a point on the boundary, multiplied by the normal at that point turns out to guarantee a descent direction for the optimisation process). The standard FEM gives a very poor approximation of the sensitivities along the boundary and so we show how to apply the Adaptive MsFEM in Chapter 4 to obtain better approximations along the boundary of the structure. Note that in this thesis we do not explore the effect of using the AMsFEM as part of the whole optimisation process but rather consider the improvement to a single step when the linear elasticity problem is solved. The main point of this chapter is to show that by using the AMsFEM a single fixed mesh can be used for the whole optimisation process and we obtain a more accurate solution along the boundary. We demonstrate this with some benchmark results in Section 5.5.

Before we begin describing the shape optimisation problem we would like to acknowledge the helpful discussions with Alicia Kim [55, 57, 56] and Peter Dunning [33, 34] who presented the problem to us and for helping to clarify the shape optimisation process. Thanks also go to Peter for providing the structure images of the benchmark problems (Figures 5-4, 5-8 and 5-12) and an initial set of shape sensitivity data using ANSYS. We would also like to acknowledge the helpful discussions with Phil Browne about shape optimisation for methods that do not use the level set approach.

5.1 Expanding the problem definition

The problem defined in Section 2.1 was restricted to very specific interface problems. This was to make the apriori convergence analysis possible. Now we can expand to more general problems that involve a bounded and coercive bilinear form. We also expand to more general mixed boundary conditions. As we will see in the later in this chapter this expansion allows us to solve linear elasticity problems with the adaptive multiscale method. Given a domain $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega$ partitioned into $\Gamma_D \cup \Gamma_N$ where $\Gamma_D \neq \emptyset$, we define the multi-dimensional spaces $\left[H^1(\Omega)\right]^d$ and $\left[H^1(\Omega)\right]^d_{0,\Gamma_D}$.

Definition 5.1. *Define the space*

$$\left[H^1(\Omega)\right]^d = H^1(\Omega) \times H^1(\Omega) \times \ldots \times H^1(\Omega) \ , \tag{5.1}$$

where $\mathbf{u}(x) \in \left[H^1(\Omega)\right]^d$ *means that* $u_i(x) \in H^1(\Omega)$ *for any* $i = 1, \ldots, d$.

Definition 5.2. *Consequently define the space*

$$\left[H^1(\Omega)\right]^d_{0,\Gamma_D} = \left\{ \mathbf{u} \in \left[H^1(\Omega)\right]^d \mid \mathbf{u} = 0 \text{ on } \Gamma_D \right\} . \tag{5.2}$$

Now suppose we are given a bilinear form $A(\cdot,\cdot)$ on $\left[H^1(\Omega)\right]^d$ that is both bounded and coercive, i.e. that

1. for any $\mathbf{u}, \mathbf{v} \in \left[H^1(\Omega)\right]^d$ then $A(\mathbf{u}, \mathbf{v}) \leq \nu_1 \|\mathbf{u}\| \|\mathbf{v}\|$,
2. for any $\mathbf{u} \in \left[H^1(\Omega)\right]^d$ then $A(\mathbf{u}, \mathbf{u}) \geq \nu_2 \|\mathbf{u}\|^2$,

for constants $\nu_1, \nu_2 > 0$ and $\|\mathbf{u}\| = \left\{ \sum_{i=1}^d \|u_i\|^2_{H^1(\Omega)} \right\}^{\frac{1}{2}}$. Then given a bounded functional $F(\cdot)$ and Dirichlet boundary data g_D on Γ_D we introduce the variational multiscale problem.

Problem 5.3. *(The Variational Multiscale Problem) Let* $\mathbf{w} \in \left[H^1(\Omega)\right]^d$ *be a function that coincides with* g_D *on* Γ_D. *Then find* $\mathbf{u} = \mathbf{u}_0 + \mathbf{w}$ *where* $\mathbf{u}_0 \in \left[H^1(\Omega)\right]^d_{0,\Gamma_D}$ *such that*

$$A(\mathbf{u}_0, \boldsymbol{\phi}) = F(\boldsymbol{\phi}) - A(\mathbf{w}, \boldsymbol{\phi}) \quad \text{for all } \boldsymbol{\phi} \in \left[H^1(\Omega)\right]^d_{0,\Gamma_D} . \tag{5.3}$$

Since $A(\cdot,\cdot)$ is bounded and coercive, and as $F(\cdot)$ is bounded then Problem 5.3 has a unique solution $\mathbf{u}_0 \in \left[H^1(\Omega)\right]^d_{0,\Gamma_D}$ because of the Lax-Milgram Theorem [20]. This gives us a very general framework to work with. An example of a bilinear form fitting the framework is

$$A(\mathbf{u}, \boldsymbol{\phi}) = \int_\Omega \nabla \mathbf{u} \cdot \mathcal{A} \nabla \boldsymbol{\phi} dx , \tag{5.4}$$

with $\mathcal{A}$ a uniformly positive definite matrix with L_∞ entries, e.g. $\mathcal{A}(x)$ being a realisation of a random field. It also allows vector valued problems to be considered, like the planar linear elasticity problem using the bilinear form

$$A(\mathbf{u}, \boldsymbol{\phi}) = \int_\Omega \left[\frac{\partial u_1}{\partial x}, \ \frac{\partial u_2}{\partial y}, \ \frac{\partial u_2}{\partial x} + \frac{\partial u_1}{\partial y} \right] \mathcal{A} \left[\frac{\partial \phi_1}{\partial x}, \ \frac{\partial \phi_2}{\partial y}, \ \frac{\partial \phi_2}{\partial x} + \frac{\partial \phi_1}{\partial y} \right]^T dx , \tag{5.5}$$

where

$$\mathcal{A}(x) = \begin{bmatrix} \lambda(x) + 2\mu(x) & \lambda(x) & 0 \\ \lambda(x) & \lambda(x) + 2\mu(x) & 0 \\ 0 & 0 & \mu(x) \end{bmatrix} , \tag{5.6}$$

and $\lambda(x), \mu(x)$ are the Lamé constants depending on the material properties at x. This example is discussed further in Chapter 5. Similarly we can also now include more general boundary conditions on $\partial\Omega$. We have already seen how to include Dirichlet conditions on Γ_D but we can also include inhomogeneous Neumann conditions on Γ_N. This is done by including it into the functional $F(\cdot)$, supposing we had a load function $\mathbf{f} \in [L_2(\Omega)]^d$ and Neumann conditions $\mathbf{g}_N \in [L_2(\Gamma_N)]^d$ then we could define $F(\phi)$ by

$$F(\boldsymbol{\phi}) = \int_\Omega \mathbf{f} \cdot \boldsymbol{\phi} \; dx + \int_{\Gamma_N} \mathbf{g}_N \cdot \boldsymbol{\phi} \; dS \; . \tag{5.7}$$

5.2 The linear elasticity formulation

In this section we will explore the mathematical definition of the linear elasticity problem which forms the setting for the topology optimisation problem which we subsequently describe. The following description fits into the framework defined in Section 5.1 and thus immediately allows the application of the Adaptive Multiscale Finite Element Method. Let $\Omega \subset \mathbb{R}^2$ and denote the displacement field of an elastic body by $\mathbf{u} \in [H^1(\Omega)]^2$ (see Definition 5.1) representing x- and y- displacements. The stress-strain relations for linear elastic materials yield the following bilinear form,

$$A_\Omega(\mathbf{u}, \mathbf{v}) = \int_\Omega \epsilon(\mathbf{u}) \cdot \mathcal{A}\epsilon(\mathbf{v}) dx \tag{5.8}$$

for any $\mathbf{u}, \mathbf{v} \in [H^1(\Omega)]^2$. In the above bilinear form $\epsilon(\mathbf{u})$ arises from the unique elements of the 2×2 infinitesimal strain tensor $\frac{1}{2}[\nabla \mathbf{u} + (\nabla \mathbf{u})^T]$ which is symmetric along with the plane stress and strain conditions (see [30] by Cook for more details), thus $\epsilon(\mathbf{u})$ is given by

$$\epsilon(\mathbf{u}) = \left[\; \tfrac{\partial \mathbf{u}_1}{\partial x}, \; \tfrac{\partial \mathbf{u}_2}{\partial y}, \; \tfrac{\partial \mathbf{u}_2}{\partial x} + \tfrac{\partial \mathbf{u}_1}{\partial y} \; \right]^T \; ,$$

and $\mathcal{A}$ is the stiffness tensor given by

$$\mathcal{A} = \begin{bmatrix} \lambda(\mathbf{x}) + 2\mu(\mathbf{x}) & \lambda(\mathbf{x}) & 0 \\ \lambda(\mathbf{x}) & \lambda(\mathbf{x}) + 2\mu(\mathbf{x}) & 0 \\ 0 & 0 & \mu(\mathbf{x}) \end{bmatrix} . \tag{5.9}$$

Note that λ and μ are the Lamé constants and are material specific at the position $\mathbf{x}$. It is more common within engineering to write these in terms of the Young's modulus E and the Poisson ratio ν of the material. So

$$\lambda(\mathbf{x}) = \frac{E(\mathbf{x})\nu(\mathbf{x})}{(1 + \nu(\mathbf{x}))(1 - 2\nu(\mathbf{x}))} \qquad \mu = \frac{E(\mathbf{x})}{2(1 + \nu(\mathbf{x}))} \; . \tag{5.10}$$

It is important to note that the bilinear form arising from planar linear eleasticity given above is both bounded and coercive. A lengthy and elementary set of calculations shows that $A(u,v) \leq \max_{x\in\Omega}(\lambda(x)+2\mu(x))\,\|u\|\,\|v\|$ and $A(u,v) \geq \min_{x\in\Omega}\{\lambda(x),2\mu(x)\}\,\|u\|^2$ since $\lambda(x),\ \mu(x) \geq 0$ for all $x \in \Omega$. Note that this shows the importance later for the fixed mesh structural optimisation problem of choosing a small but non-zero material coefficient for the ghost material. We will explore this again later when we discuss the fixed mesh problem.

Therefore the linear elasticity problem can be stated as a specific instance of the variational multiscale problem (see Problem 5.3) since the Lamé constants can incorporate many scales and many different materials. We will discuss the specifics of the boundary conditions that arise in structural optimisation problems later, but first state the general linear elasticity deformation problem. For now, suppose $\mathbf{u}$ coincides on $\Gamma_D \subset \partial\Omega$ with a fixed displacement function $\mathbf{g}$, suppose also that Ω is subject to a body force $\mathbf{b}$ and traction boundary conditions $\mathbf{t}$ on $\Gamma_N \subset \partial\Omega$. If $\mathbf{w} \in [H^1(\Omega)]^2$ is any extension of $\mathbf{g}$ into $[H^1(\Omega)]^2$, then the general form of our problem is:

Problem 5.4. *Find* $\mathbf{u} = \mathbf{u}_0 + \mathbf{w}$ *where* $\mathbf{u}_0 \in [H^1_{0,\Gamma_D}(\Omega)]^2$ *such that*

$$A_\Omega(\mathbf{u}_0, \boldsymbol{\phi}) = \int_\Omega \mathbf{b}\cdot\boldsymbol{\phi}dx + \int_{\Gamma_N} \mathbf{t}\cdot\boldsymbol{\phi}dS - A_\Omega(\mathbf{w},\boldsymbol{\phi}) \quad \textit{for all } \boldsymbol{\phi} \in [H^1_{0,\Gamma_D}(\Omega)]^2\ . \quad (5.11)$$

To make the following section on the structural optimisation process simpler to convey we make the following assumption.

Assumption 5.5. *We restrict to the case when there is no body force, i.e.* $\mathbf{b} = 0$.

Instead of considering body forces such as gravity acting on the design structure, we consider only fixed displacements and surface tractions acting on the boundary of the structure. Firstly the fixed displacements are given by $\mathbf{g}(x)$ on the Dirichlet boundary Γ_D. Typically Γ_D defines the fixed points of the structure and so in this chapter we will set $\mathbf{g}(x) = 0$.

Most applications in structural optimisation predominantly consider traction forces. The optimisation process will change the boundary of the structure which means that it will be necessary to divide the boundary into parts that are allowed to change and parts that are not. This requires some post processing to ensure that Γ_D does not shrink to the empty set as this would permit rigid body motions and thus lose uniqueness of the solution of the linear elasticity problem. The other requirement to consider is that the points where a traction boundary force is specified must not change either, to this end we split Γ_N into two parts $\Gamma_0 \cup \Gamma_t$. Γ_t is the part of Γ_N experiencing a specified

traction force and Γ_0 is the part of the boundary that is allowed to move and to change. Therefore we let $\mathbf{t} = 0$ on Γ_0.

On Γ_t we define two classes of loading tractions. The first is a point loading traction where a specified force f is applied in direction $\boldsymbol{\nu}(\mathbf{x})$ to a particular point $\mathbf{x}_0 \in \Gamma_t$, thus

$$\mathbf{t}(\mathbf{x}) = f\delta(\mathbf{x} - \mathbf{x}_0)\boldsymbol{\nu}(\mathbf{x}) \ . \tag{5.12}$$

The second class is an area traction where a specified force per unit area, f_{unit}, in direction $\boldsymbol{\nu}$ is applied along all or part of Γ_t, given by

$$\mathbf{t} = f_{\text{unit}}\boldsymbol{\nu} \ . \tag{5.13}$$

This gives a broad framework that includes many linear elasticity deformation problems. We will explore three such problems in Section 5.5 when we apply the new adaptive multiscale method to structural optimisation. However next we explore the structural optimisation process.

5.3 The structural optimisation problem

While we will formulate the application of the Adaptive Multiscale method to general linear elasticity problems, the motivation for this chapter comes from an engineering problem. Mechanical engineers often seek to design the stiffest structure under a variety of loading conditions. Conventionally this is done by hand where the engineer creates an initial design, it is then simulated to assess its performance and then re-designed to improve the structure. This cycle of design, analysis and re-design is normally all done by hand and takes a large amount of time. An emerging field in mechanical engineering is that of structural optimisation where the aim is to automate this design cycle. This is done by solving the following constrained minimisation problem over the set of admissible shapes $\mathcal{U}_{\text{ad}} = \left\{\Omega_S \subset \mathbb{R}^2 \text{ is a connected domain such that } \Gamma_D \cup \Gamma_N \subset \partial\Omega_S\right\}$.

Problem 5.6. *Find a domain $\Omega_S^* \in \mathcal{U}_{ad}$ such that an objective function J (for example the compliance objective function J_C in (5.20) later) is minimised subject to the static equilibrium equations and restricted material. Therefore,*

$$J(\mathbf{u}, \Omega_S^*) = \min_{\Omega_S \in \mathcal{U}_{ad}} J(\mathbf{u}, \Omega_S) \tag{5.14}$$

subject to

$$A_{\Omega_S}(\mathbf{u}, \boldsymbol{\phi}) = \int_{\Omega_S} \mathbf{b} \cdot \boldsymbol{\phi} dx \quad \textit{for all } \boldsymbol{\phi} \in [H^1_{0,\Gamma_D}(\Omega_S)]^2 \textit{ , (equilibrium equations)} \tag{5.15}$$

$$\mathbf{u} = \mathbf{g} \quad \textit{on } \Gamma_D \textit{ ,} \qquad (\mathcal{A}\epsilon(\mathbf{u})) \cdot \boldsymbol{\nu}(\mathbf{x}) = \mathbf{t}(\mathbf{x}) \textit{ for } \mathbf{x} \in \Gamma_N \textit{ , (boundary conditions)} \tag{5.16}$$

$$\int_{\Omega_S} dx \leq \gamma^* \textit{(material volume constraint)} \tag{5.17}$$

where $\mathbf{u} \in [H^1(\Omega_S)]^2$ *is the displacement field found from solving (5.15) and (5.16). Also* γ^* *is a fixed maximum volume.*

Instead of searching over all structures in $\mathbb{R}^2$ the problem is usually limited to a single fixed design domain Ω that bounds the region containing the boundary conditions. There are a wide range of ideas for how to perform this automated design cycle (see Section 1.2.4) but for this thesis we consider one in particular, the level set approach to structural optimisation. The idea behind this approach is to avoid an explicit description of the structure and its boundaries but rather define it implicitly by a level set function defined on the design domain, Ω given by

$$\mathcal{L}(x) = \begin{cases} < 0 & \text{if} \quad x \in \Omega_S \\ = 0 & \text{if} \quad x \in \Gamma_S := \partial\Omega_S \\ > 0 & \text{if} \quad x \in \Omega \backslash \overline{\Omega_S} \end{cases} , \tag{5.18}$$

where Ω_S is the domain of the structure and Γ_S is the boundary of the structure. Allaire et al [10] proposed updating the implicit shape function over time (or rather design iterations) by finding a normal velocity $V(x)$ (which we will state below) and iteratively solving a Hamilton-Jacobi type formulation,

$$\frac{\partial \mathcal{L}(x,t)}{\partial t} + \nabla \mathcal{L}(x,t) \cdot \frac{dx}{dt} = \frac{\partial \mathcal{L}(x,t)}{\partial t} + \nabla \mathcal{L}(x,t) \cdot (V(x)\mathbf{n}) = 0$$

where $\mathbf{n} = \nabla \mathcal{L}(x,t) / |\nabla \mathcal{L}(x,t)|$ is the normal direction at x. This is then discretised and written as an update scheme given by

$$\mathcal{L}^{k+1}(x) = \mathcal{L}^k(x) - \Delta t \left| \nabla \mathcal{L}^k(x) \right| V(x) \qquad \text{for any } x \in \Gamma_S \text{ ,} \tag{5.19}$$

where k is the iterative step, Δt is the discrete time step and V is the the speed normal to the boundary of the structure. The time step is limited by the Courant-Freidrichs-Lewy (CFL) condition and thus the key to optimising the shape function is the normal component of the velocity, V.

The objective of the structural optimisation process is to then minimise a desired objective function, $J(u, \Omega_S)$, with respect to the set of admissible shapes $\mathcal{U}_{\text{ad}}$. This can

be stated in a very general form but for this thesis we consider a specific objective, the compliance of the structure. The compliance is a measure of the strain energy within the structure when it is subjected to loading forces and relates to the stiffness of the structure. The compliance function is given by

$$J_C(\mathbf{u}, \Omega_S) = \int_{\Omega_S} \epsilon(\mathbf{u}) \mathcal{A} \epsilon(\mathbf{u}) \ , \tag{5.20}$$

and in order to minimise the strain energy we consider the 'shape derivative' of the current structure. The shape derivative was shown by Allaire to take the form

$$J'_C(\mathbf{u}, \Omega_S) = \int_{\Gamma_0} (\epsilon(\mathbf{u}) \mathcal{A} \epsilon(\mathbf{u}))\, V \ , \tag{5.21}$$

where Γ_0 is the free boundary part of Γ_S and V is the normal component of the velocity as in (5.19). With the aim of minimising the compliance function we define the shape sensitivity along the free boundary by

$$\zeta(\mathbf{u}) := \epsilon(\mathbf{u}) \mathcal{A} \epsilon(\mathbf{u}) \ . \tag{5.22}$$

Thus to minimise the compliance $J_C(\mathbf{u}, \Omega_S)$ we let

$$V = -\zeta(\mathbf{u}) \ , \tag{5.23}$$

such that

$$J'_C(\mathbf{u}, \Omega_S) = -\int_{\Gamma_0} |\zeta(\mathbf{u})|^2 < 0 \ ,$$

thereby ensuring that the compliance function is decreasing. This is not a complete description however, since the above description does not include the constraint of limited material. We do not explore the specifics in this thesis but simply note that the constrained problem is converted into an unconstrained problem with a Lagrange multiplier. The compliance function to be minimised becomes

$$\widetilde{J}_C(\mathbf{u}, \Omega_S) = J_C(\mathbf{u}, \Omega_S) + \lambda \int_{\Omega_S} 1 \ ,$$

and $V = \lambda - \zeta(\mathbf{u})$. The technicalities of the choice of λ are dealt with in the engineering literature but we can see that the accuracy of V depends on the accuracy of the shape sensitivity $\zeta(\mathbf{u})$ along the boundary. We give an example in order to set the topology optimisation process in context.

The first example that we explore is the 2D short cantilever, a structure fixed along its left hand edge and subjected to a downward point load at the center of the right hand side. In the framework we have outlined so far, we initially start with a solid bar with

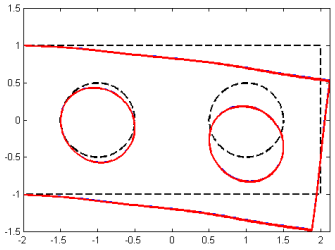
two holes in it. Let

$$\Omega_S = ([-2,2]\times[-1,1]) \setminus \left\{x \in \mathbb{R}^2 \mid \|x \pm 1/2\|_2 < 1/2\right\} \ .$$

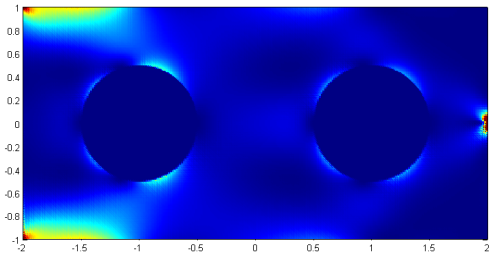
For the boundary conditions let $\Gamma_D = \{x \in \Omega_S \mid x_1 = -2\}$ with $\mathbf{g}(x) = 0$ and $\Gamma_N = \partial\Omega_S \backslash \Gamma_D$. Within Γ_N let $\Gamma_t = (2,0)^T$ and $\mathbf{t} = \delta(\mathbf{x} - \Gamma_t)(0,-1)^T$, i.e. the cantilever experiences a downward force of $(0,-1)^T$ at the point $(2,0)^T$ and zero traction on the rest of Γ_N.

The original starting structure Ω_S is shown as the dashed line in Figure 5-1(a) and the structure after deformation is given by the solid red line. Figure 5-1(b) shows the shape sensitivity going from low (dark blue) to high (red), in particular note the distribution along the edges of the two circles. The resulting velocity field, V, adds more material where the shape sensitivity is higher (light blue areas along the circle) and removes material where is lower (the dark blue areas along the circle). The same process occurs along the outer boundary with the restriction that it cannot move at $(2,0)^T$ and cannot detach from the edge where $x = -2$. The resulting structure after 200 steps of the optimisation process is shown in Figure 5-1(c) where most of the initial material is removed, as can be seen the aim is to have an even distribution of strain across the structure. Over the course of the optimisation process the key structure emerges where the loading point on the right is connected by two diagonal supports to the two points on the left that are furthest apart, the microstructure in the middle reduces the size of the main supports. The 2D cantilever problem given here is a common example in the engineering literature and for example is given in [10] where they perform 100 optimisation steps. In this thesis we do not consider the optimality of this structure or technicalities due to new hole insertion into the structure but note that both of these issues are a major concern within the topology optimisation field. In this thesis we are instead focusing on accuracy of the solution to the linear elasticity problem, a single step of the optimisation process, and its effect on the shape sensitivity.

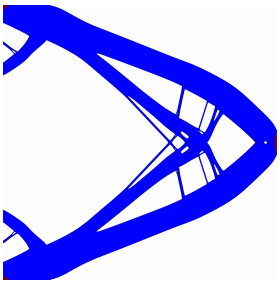
So far we have outlined the general shape optimisation problem, one of the major difficulties lies in the details of how the equilibrium equations are solved. Most shape optimisation methods solve this stage using finite element analysis, however conventional methods require that a mesh be fitted to the domain Ω_S to obtain good accuracy. This is very expensive for shape optimisation because the finite element mesh has to be reformed at each step as the level set function $\mathcal{L}$ changes. Instead a single fixed mesh is used and the voids, $\Omega \backslash \Omega_S$, are filled with a weak 'ersatz' material ([10],[78]). This



(a) 2D cantilever under deformation (10x scale factor for the deformation).



(b) Shape sensitivity across the cantilever.



(c) An example of an optimised 2D cantilever (Courtesy of P.A. Browne).

Figure 5-1: *Diagram showing a 2D cantilever under deformation and the corresponding shape sensitivity distribution. The shape sensitivity scales from dark blue to red corresponding to areas of low strain to high strain and was calculated using standard finite elements on a* 256×128 *fine mesh.*

creates a discontinuous Young's modulus given by

$$E(\mathbf{x}) = \begin{cases} E_{\text{material}} & x \in \Omega_S \\ E_{\text{ersatz}} & \text{otherwise} \end{cases}$$

but the Poisson ratios are equal. Typically E_{ersatz} is taken to be a fraction of E_{material}, $E_{\text{material}}/10^5$ for example.

Defining the ersatz material allows the velocity function V (5.23) to be defined throughout the design domain Ω, and not just within the structure Ω_S. Therefore any update process that uses the fixed mesh to update the implicit shape function is still defined within the voids. Many engineering applications base the level set function $\mathcal{L}$ on the fixed mesh using interpolation between nodes. This is done for speed of execution but loses the key idea that the level set function is completely separate from the method used for solving the linear elasticity problem.

The problem with fixed grid methods that use the standard finite element method for

solving the linear elasticity interface problem is that the shape sensitivty along the interface is not smooth. As we have seen in Chapter 2 the error near the interface depends on the size of the high coefficient part of a cut element to the size of the whole element. We combined these together in our estimate to obtain

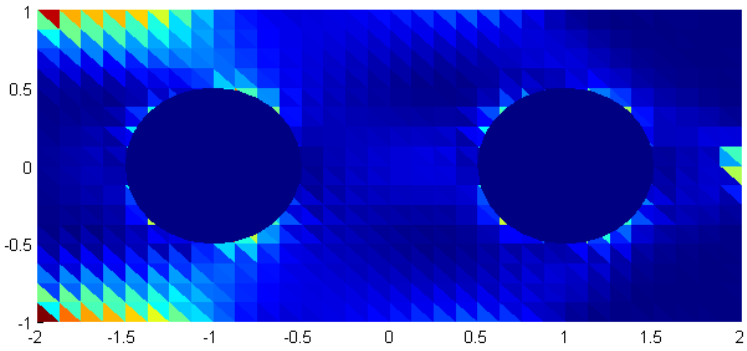
$$\eta_H = \max_{\tau \in \mathcal{T}_H^C(\Omega)} H_\tau / \rho_{K(\tau)} \ ,$$

but the problem is slightly worse in the case of structural optimisation. Experimental results show that the optimisation process struggles to converge when two neighbouring elements have very different values of $H_\tau/\rho_{K(\tau)}$, i.e. one element cuts Ω_S by a little and the other neighbouring elements cut Ω_S by a lot. This means the error in two neighbouring elements may be very different and thus the velocity V is very different too. What occurs is an undesired roughness in the shape function $\mathcal{L}(x)$ along the edges which causes poor convergence ([87], [86]). The varying errors present themselves by sharp changes in the values of the shape sensitivity $\zeta(u)$ along the interfaces. What is desired is a smooth, and accurate, sensitivity profile along the interface obtain by averaging the shape sensitivity across element edges.

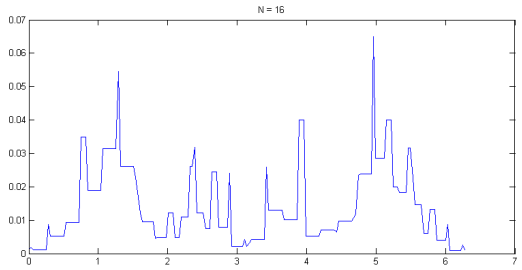
Introduction of the ersatz material produces discontinuous Lamé constants (5.10) and a discontinuous material matrix $\mathcal{A}$ (5.9). Thus the linear elasticity problem becomes an instance of a high contrast elliptic interface problem similar to the interface problem in Chapters 2 and 3. As E_{ersatz} tends to zero the problem becomes more realistic but as we have seen in Chapters 2 and 3 this causes the solution to blow up as the ellipticity is lost. This allows us to implement the Adpative MsFEM in Chapter 4 to provide a method that both uses a fixed mesh and adapts the basis functions to give a more accurate result along the boundary of the structure.

We demonstrate the large jumps in the shape sensitivity for the standard FEM by solving the linear elasticity problem for the 2D cantilever example shown in Figure 5-1(a). Figure 5-2(a) shows the shape sensitivity distribution based on a 32×16 uniform mesh (note that the shape sensitivity is constant in each element and each material for linear hat functions). The shape sensitivity around the boundary of the left hand circle is then shown in Figure 5-2(b) and the equivalent graph for a 256×128 uniform mesh is shown in Figure 5-2(c) for comparison. Although not completely smooth, it is smoother than Figure 5-2(b). This shows the difficulty with the optimisation process, for speed the coarse 32×16 mesh is desired but then the 256×128 mesh is required for a good solution where the optimisation process converges quickly.

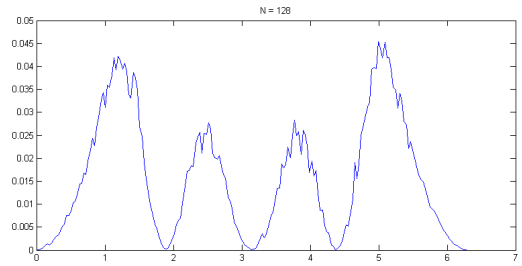
Typically it is not possible to have a fixed fine mesh that sufficiently resolves the boundary of the structure to give a smooth shape sensitivity, thus profiles as in Figure



(a) Shape sensitivity for a 32×16 uniform mesh with $\mathbb{P}_1$ finite elements.



(b) Shape sensitivity around the boundary of the left hand circle for a 32×16 uniform mesh.



(c) Shape sensitivity around the boundary of the left hand circle for a 256×128 uniform mesh.

Figure 5-2: *Shape sensitivity comparison between a* 32×16 *uniform mesh and a* 256×128 *uniform mesh along the left hand circle. Note the extreme jumps in the sensitivity for the coarser mesh.*

5-2(c) are extremely difficult to obtain.

Wei, Wang and Xing show the poor convergence of the standard FEM (referred to as the density method) in [87] and apply X-FEM to obtain a more accurate shape sensitivity distribution. X-FEM follows a similar idea of using better basis functions to model the discontinuity in strain energy but does so by adding additional degrees of freedom and is still confined to basis functions that are the product of a polynomial and some predefined choice of enrichment functions (e.g. the Heaviside step function). We refer back to Section 1.2.2 for more discussion on the extended finite element method.

Other heuristic engineering approaches include approximating the shape sensitivity ζ (see (5.22)) at the nodes by an area weighted average

$$\zeta_i = \left(\sum_{\substack{\tau \in \mathcal{T}_H(\Omega) \\ n_i \in \tau}} \zeta|_\tau (n_i) \int_{\tau \cap \Omega_S} dx \right) / \left(\sum_{\substack{\tau \in \mathcal{T}_H(\Omega) \\ n_i \in \tau}} \int_{\tau \cap \Omega_S} dx \right) ,$$

where $n_i \in \mathcal{N}(\mathcal{T}_H(\Omega))$ and linear interpolation is used between nodes. Note that the optimisation process seeks a structure such that $\zeta(n_i) = 0$ for $n_i \in \Gamma_0$ but $\zeta(x)$ can be defined everywhere. This area weighted approach is motivated by analytical results for a few model cases where the interface cuts the element as a straight line and all depended on the relative proportions of material in neighbouring elements. The area weighted approach is still unsatisfactory as García-Ruíz and Steven show in [40] that the maximum stress error is still large. Instead they suggest a least squares approach to fit a polynomial to data points within a certain radius.

The least squares approach seeks to obtain more accurate values for the shape sensitivity at the nodes by fitting a polynomial to the shape sensitivity at the Gauss points of elements that intersect Ω_S within a certain radius r_{ls} (e.g. $r_{ls} = 2H$). Once this polynomial is found we can then calculate more accurate values of the sensitivity along the boundary $\partial\Omega_S$. This least squares approach offers an improvement over the weighted average but introduces additional inaccuracies. If a structure contains two interfaces that are close together (their separation distance is less than the least squares radius r_{ls}) as in the interior corners of the bridge structure in Section 5.5.2 then the shape sensitivity from one edge will influence that on the other edge. For example in Figure 5-8 at the point $(4, 4)$ the horizontal edge between $(4, 4)$ and $(12, 4)$ influences the shape sensitivity along the diagonal edge from $(4, 4)$ to $(12, 14)$ under the least squares scheme.

These heuristic techniques for improving the smoothness of the shape sensitivity stem from the fact that the standard FEM gives a poor and uneven error along the interface depending on how much of the structure intersects each element. It is unknown how smoothing the poor solution resulting from the standard FEM affects the structural optimisation process. Instead, in this chapter, we avoid the weighted average and least squares approaches (with are essentially smoothing procedures) and return to the original idea by Allaire and take simple nodal averages,

$$\zeta(x) = \left(\sum_{\substack{\tau \in \mathcal{T}_H(\Omega) \\ x \in \tau}} \zeta|_\tau (x) \right) / \left(\sum_{\substack{\tau \in \mathcal{T}_H(\Omega) \\ x \in \tau}} 1 \right) ,$$

for $x \in \partial\Omega_S$ but with a more accurate solution along the edges. This is where the Adaptive Multiscale Finite Element Method can be applied, as we will see in the following section.

5.4 AMsFEM applied to structural optimization

The problems with rapidly varying shape sensitivities arise when a finite element mesh does not sufficiently resolve the interfaces. We can rectify this by applying the adaptive multiscale finite element method introduced in Chapter 4. Thus we introduce multiscale basis functions which provide a smoother shape sensitivity whilst still only needing a coarse finite element mesh. The method follows the same framework outlined in Algorithm 2 but with some technical changes due to the increased dimensionality of the problem.

Since the displacement field $\mathbf{u}$ is a 2-dimensional vector it is necessary to define a multidimensional set of basis functions. Normally we define a set of D-dimensional basis functions $\mathbf{\Phi}_{i,j}$ $i = 1, ..., N,\ j = 1, ..., D$ using the previous scalar linear hat functions in Chapter 2. For example in 2D for a 2D displacement field we denote

$$\mathbf{\Phi}_{i,1} = \begin{pmatrix} \phi_i \\ 0 \end{pmatrix}, \quad \mathbf{\Phi}_{i,2} = \begin{pmatrix} 0 \\ \phi_i \end{pmatrix}$$

for each $n_i \in \mathcal{N}(\mathcal{T}_H(\Omega))$ and ϕ_i is the usual linear hat function. Instead for the multiscale basis functions we will solve a local problem as in the adaptive multiscale method in Chapter 4 to obtain the basis

$$\left\{\mathbf{\Phi}^{MS}_{1,1}, ..., \mathbf{\Phi}^{MS}_{1,D}, \mathbf{\Phi}^{MS}_{2,1}, ..., \mathbf{\Phi}^{MS}_{2,D}, ..., \mathbf{\Phi}^{MS}_{N,1}, ..., \mathbf{\Phi}^{MS}_{N,D}\right\}$$

where D is the dimension of the solution ($D = 2$ for the planar linear elasticity problem) and N is the number of coarse mesh nodes. Note that each $\mathbf{\Phi}$ is itself a D-dimensional vector.

The next detail involves modifying the boundary condition described in stage 1 of the iterative step (Algorithm 1). Let τ be an element currently being processed and $\tilde{\tau}$ its corresponding extension, then we can use Definition 4.7 to define a multi-dimensional version of the boundary condition for the local extended basis functions $\mathbf{\Psi}^{MS}_{\tilde{\tau},\cdot,\cdot}$. This was defined for a scalar valued basis function in Chapter 4 which can be easily extended using the following formulation. Let

$$\left(\mathbf{\Psi}^{\mathrm{MS}}_{\tilde{\tau},j,k}\right)_l\Big|_{e_i} = \delta_{lk}\mathcal{P}_{j,e_i}\mathbf{u}_k \tag{5.24}$$

for $i, j = 1, 2, 3$ and $k, l = 1, ..., D$. This creates a set of basis functions that are non-zero in only one of the D dimensions. The iterative step then proceeds to solve the same local homogeneous problem (4.6) to obtain the extended basis functions $\left\{\mathbf{\Psi}^{\mathrm{MS}}_{\tilde{\tau},j,k} \,|\, j = 1, 2, 3 \quad k = 1, ..., D\right\}$.

The method by which the extended basis functions on $\tilde{\tau}$ are combined on τ to obtain $\mathbf{\Phi}^{\mathrm{MS}}_{\tau,\cdot,\cdot}$ also requires an extension to multiple dimensions. The extended basis functions are combined in exactly the same way as in Section 4.3.4 to obtain D matrix systems of the form (4.9) given by

$$C_l\Psi = [c_{ji,l}]\left[\left(\mathbf{\Psi}^{\mathrm{MS}}_{\tilde{\tau},i,l}(n_k)\right)_l\right] = \delta_{jk} = I_3 \ , \tag{5.25}$$

where $i,j,k = 1,...,3$ and $l = 1,...,D$. We then use these coefficients $c_{ji,l}$ to calculate the new nodal basis functions $\mathbf{\Phi}^{\mathrm{MS}}_{\tau,i,l}$ on τ just as in (4.7) but with $i,j = 1,...,3$ and $l = 1,...,D$. Therefore the basis functions are given by

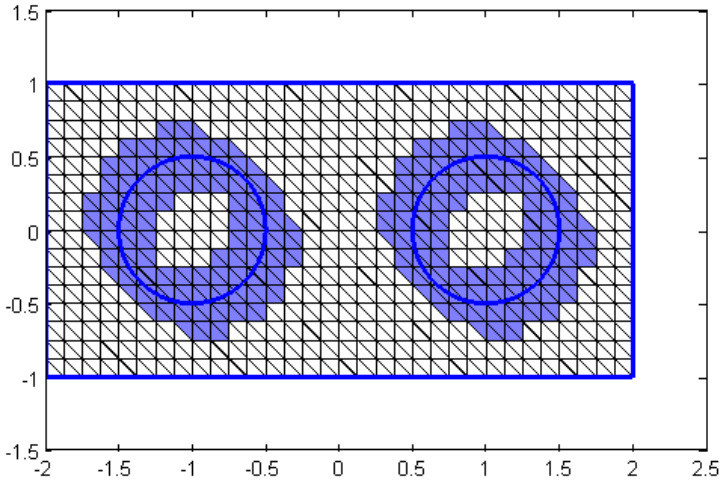
$$\left(\Phi^{\mathrm{MS}}_{\tau,j,l}\right)_l(x) = \sum_{i=1}^{3} c_{ji,l}\left(\mathbf{\Psi}^{\mathrm{MS}}_{\tilde{\tau},i,l}\right)_l(x) \ , \left(\Phi^{\mathrm{MS}}_{\tau,j,l}\right)_k(x) = 0 \tag{5.26}$$

for $k \neq l$ and where $i,j = 1,2,3$ with $k,l = 1,...,D$. We then use Definition 4.9 to define global basis functions and Notation 2.45 to define the local-to-global mapping. This produces the non-conforming basis functions $\mathbf{\Phi}^{\mathrm{MS}}_{i,j}$ for $i = 1,...,N$ $j = 1,...,D$. At this point the enhanced ALG-MsFEM averages the edges of the basis functions to make a conforming method. Thus if two elements τ and τ' share an edge e then for any basis function $\mathbf{\Phi}^{\mathrm{MS}}_{i,j}$ with $i = 1,...,D$ $j = 1,...,D$ the new conforming basis function takes the values

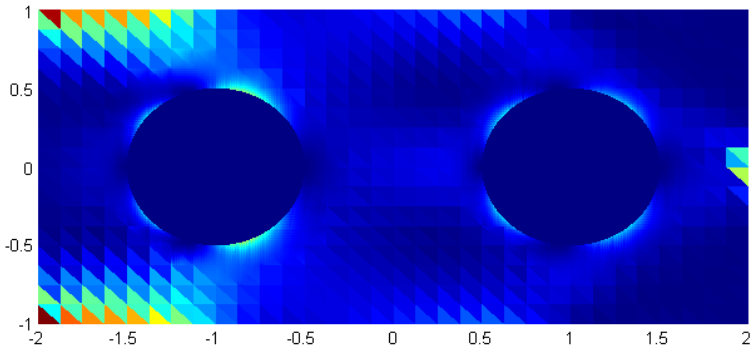
$$\frac{\mathbf{\Phi}^{\mathrm{MS}}_{i,j}\Big|_{\tau} + \mathbf{\Phi}^{\mathrm{MS}}_{i,j}\Big|_{\tau'}}{2} \ .$$

The process then proceeds as in Algorithm 2 with a solve of the global linear elasticity problem and then iterated until convergence.

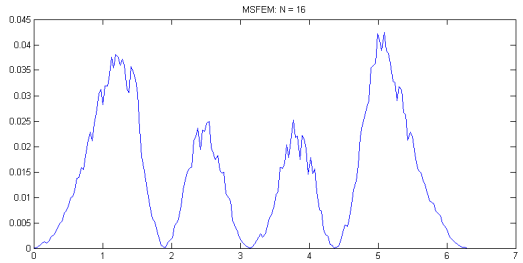
To see how well the adaptive multiscale finite element method performs for improving the shape sensitivity function we return to the 2D cantilever example. Figure 5-3(a) shows the mesh used and the shaded elements represent those for which the multiscale basis functions are used, all other elements use $\mathbb{P}_1$ linear elements. The coarse mesh uses a mesh diameter of $H = 1/8$ and a subgrid element diameter of $h = H/8$. Figure 5-3(b) shows the resulting shape sensitivity distribution which is much smoother along the interfaces than the standard FEM in Figure 5-2(a). To confirm this Figure 5-3(c) displays the corresponding graph of shape sensitivity for the left hand inclusion plotted as a function of polar angle starting at $(-0.5, 0)$. Note that this should be compared to the standard FEM graph in Figure 5-2(b) and the fine scale solution in Figure 5-2(c) also reproduced in Figure 5-3(d).



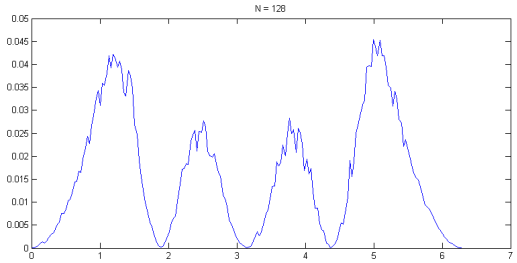
(a) The 32×16 uniform mesh for the AMsFEM with multiscale basis functions in shaded elements.



(b) Shape sensitivity for a 32×16 uniform mesh with multiscale finite elements.



(c) Shape sensitivity along the left hand circle for a 32×16 uniform mesh using the Adaptive MsFEM.



(d) Shape sensitivity along the left hand circle for a 256×128 uniform mesh with standard FEM.

Figure 5-3: *Shape sensitivity comparison between a* 32×16 *uniform mesh with multiscale basis functions and a* 256×128 *uniform mesh with standard FEM along the left hand circle. Note the lack of extreme jumps compared to Figure 5-2(b).*

5.5 Benchmark problems

To see the improvement that the Adaptive MsFEM offers for the linear elasticity problem that arises in topology optimisation we explore three benchmark examples. For each we will examine the shape sensitivity distribution along part of the boundary of the structure, $\Gamma \subset \partial\Omega_S$, to show that a much smoother result is obtained using multiscale basis functions. This section is designed to demonstrate the capability of the adaptive method to provide better sensitivities but we do not apply the method for the entire optimisation process. Further work could prove the viability of the method for optimisation and investigate any improvements in convergence to an optimal solution. The main drawback of the adaptive method being speed of execution, however, we will show in Chapter 6 that the adaptive method is very scalable on a parallel cluster.

5.5.1 A Hole in a plate

The first example we consider is a common engineering example where a square plate with a circular hole is stretched horizontally. Since the problem is symmetric about the x- and y- axis we can reduce the problem to just a quarter plate with no displacement in the y-direction on the x-axis and no displacement in the x-direction on the y-axis. Mathematically we have the design domain $\Omega = [0, 30]^2$ and the plate $\Omega_S = [0, 30]^2 \setminus \{x \in \Omega \mid \|x\|_2 < r_0\}$ where $r_0 = 15$. The displacement field $\mathbf{u}$ is subject to the boundary conditions

$$\mathbf{u}_1(\mathbf{x}) = 0 \qquad \text{if} \quad \mathbf{x}_2 = 0 \qquad\qquad\qquad \mathbf{u}_2(\mathbf{x}) = 0 \qquad \text{if} \quad \mathbf{x}_1 = 0$$

and zero traction on the remaining boundary. The tension applied along the right hand edge ($\mathbf{x}_1 = 30$) is $f_{\text{unit}} = 1$ in direction $\boldsymbol{\nu} = (1, 0)^T$ (recall Section 5.2 where this refers to the force per unit area). These loading conditions are shown in Figure 5-4(a).

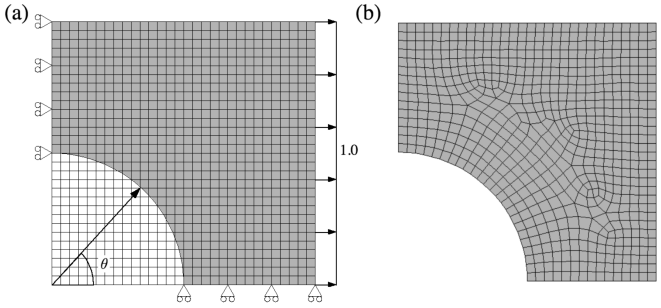


Figure 5-4: *The mesh and loading layout for the hole in a plate problem.*

For this experiment we take a Young's modulus $E = 1$ for the material and $E = 10^{-5}$

for the ersatz material $\{x \in \Omega \mid \|x\|_2 < r_0\}$, both with a Poisson ratio $\nu = 0.3$.

We know from Chapter 2 that if the mesh resolves the interface, the boundary of the quarter hole in this case, then we obtain optimal convergence. We compare the strain energy distributions obtained from various fixed mesh methods with those obtained using a fitted mesh (Figure 5-4(b)). The absolute size of the velocity is not important in topology optimisation but simply the relative speed between different parts of the interface, and for this reason the shape sensitivity is normalised along the interface. Figure 5-5 shows the shape sensitivity distribution for the fitted mesh solution with simple nodal averaging.

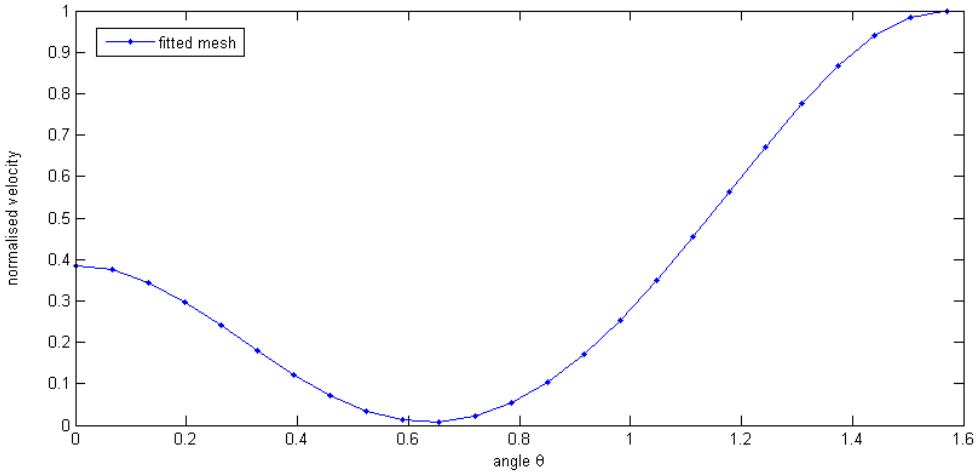
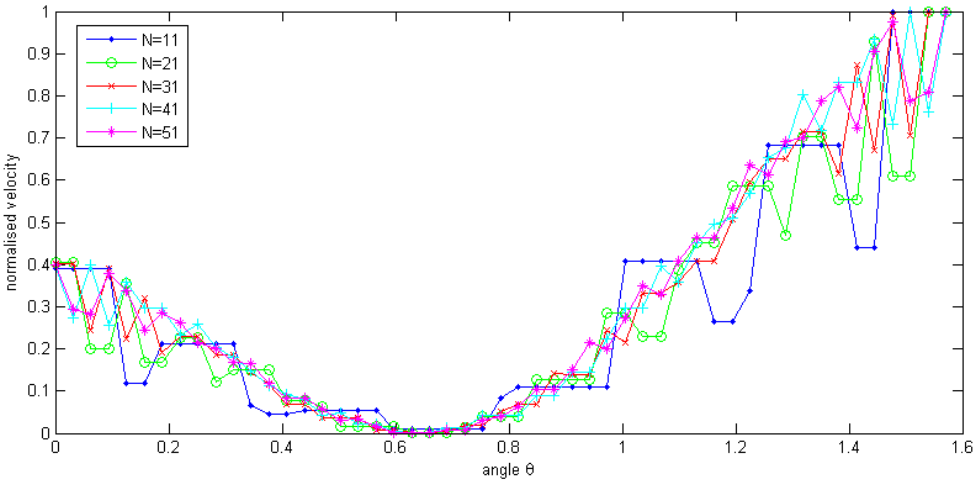


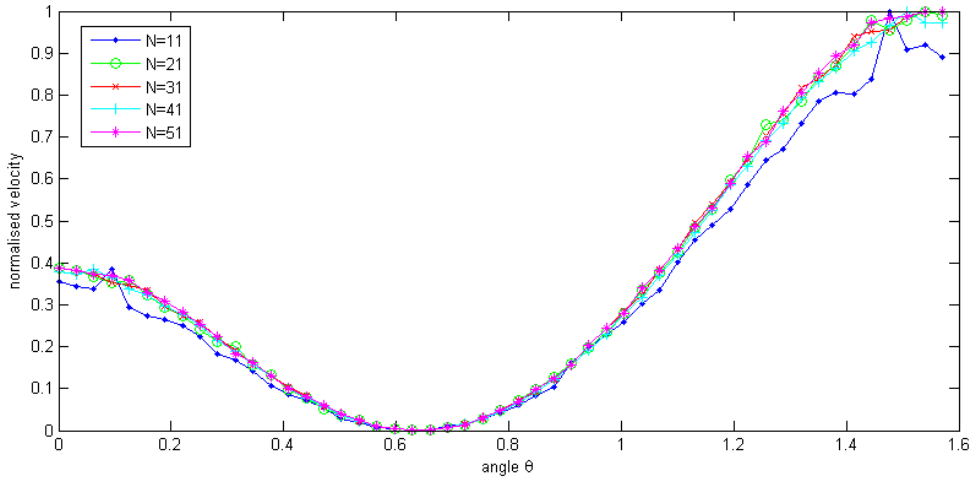
Figure 5-5: *The shape sensitivity along the circular hole using a fitted mesh measured at the nodes on the interface.*

The fitted mesh solution is calculated using the engineering software package ANSYS with bilinear quadrilateral (Q1 in the maths literature) finite elements and shape sensitivities calculated by post processing. The fitted mesh results then allow a comparison to the fixed grid standard FEM results in Figure 5-6(a) where triangular linear elements were used. Solutions to the hole in a plate problem were calculated for several mesh sizes $N = 11, 21, 31, 41, 51$ where $H = 30/N$. The same experiments were run using AMsFEM with a subgrid size of $h = H/8$, 5 iterations and applying multiscale basis functions in a 2-element band around the interface. These results are shown in Figure 5-6(b).

Note that since N is odd the ends of the circular hole ($\theta = 0, \pi/2$) fell in the middle of the element. An dditional difficulty occurs for even N as the hole becomes tangential to the elements at these two end points. We see that the AMsFEM produces consistently smoother and more accurate shape sensitivity profiles along the interface even when



(a) Standard FEM.

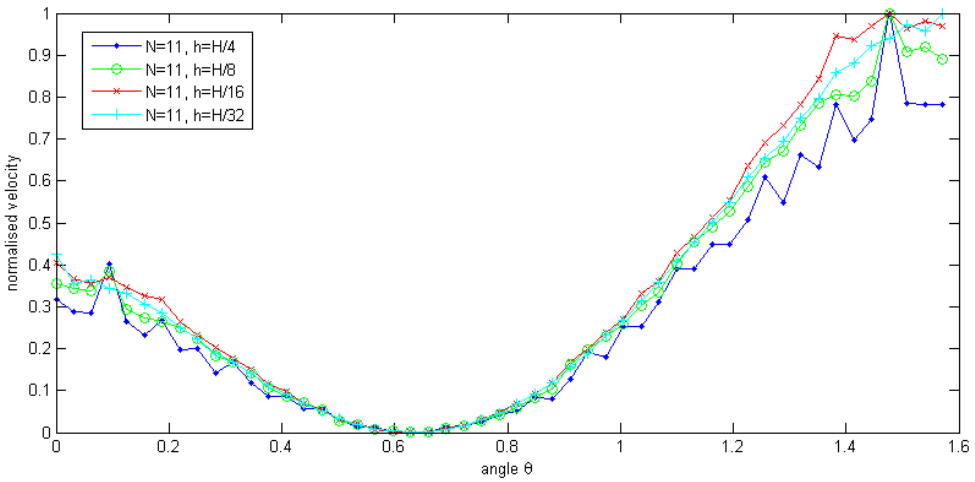


(b) AMsFEM.

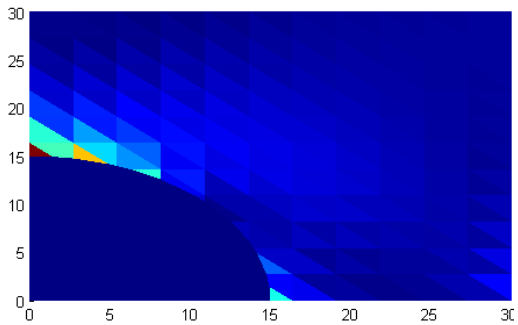
Figure 5-6: *The shape sensitivity along the circular hole using a fixed uniform mesh for both the standard FEM and AMsFEM and for varying mesh diameters where* $H = 30/N$.

the mesh H becomes small. For $N = 51$ the standard FEM still has large jumps around the $\theta = 0, \pi/2$ areas which are only mild in the AMsFEM solution for $N = 11$. This can be further improved for such a coarse mesh by decreasing the subgrid h used in each local problem. This is shown in Figure 5-7(a) where we consider a fixed uniform mesh with $H = 30/11$ and vary h with $H/4$,$H/8$,$H/16$ and $H/32$ (compare to the standard FEM with $N = 11$ given by the blue line with dots in Figure 5-6(a)). Decreasing h produces a smoother profile but also introduces the idea that the subgrid size h can be decreased or increased with each optimisation step if the level set becomes more or

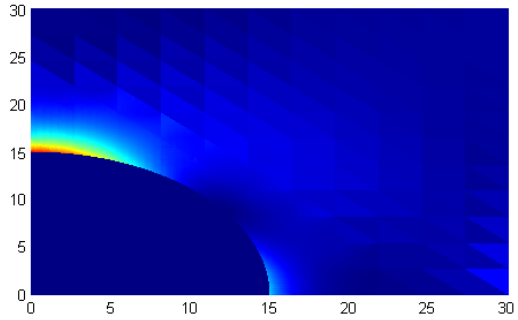
less complicated. This adds another level of control without losing the fixed (possibly uniform) coarse mesh.



(a) Shape sensitivity for varying subgrid h.



(b) Shape sensitivity for the standard FEM with an 11×11 mesh.



(c) Shape sensitivity for the AMsFEM with an 11×11 mesh and subgrid $h = H/32$.

Figure 5-7: *The shape sensitivity along the circular hole using a fixed* 11×11 *uniform mesh for both the standard FEM and AMsFEM and with varying subgrid h for the AMsFEM.*

5.5.2 Bridge problem

The second benchmark example that we explore is a bridge structure under vertical loading as seen in Figure 5-8(a). The design structure Ω_S is contained within the domain $\Omega = [0, 28] \times [0, 18]$ and the structure specifications can be obtained from Figure 5-8(a) noting that each square has sides of unit length. The bridge is fixed $(\mathbf{u}(x) = 0)$ at $x = (0, 0)^T$ and fixed in the y-direction $(\mathbf{u}_2(x) = 0)$ at $x = (28, 0)^T$. Traction free boundary conditions are applied to the rest of the boundary except along

the top where a vertical downwards force is applied. Along $\{x \in \partial\Omega_S \mid x_2 = 18\}$ we apply the unit traction $f_{\text{unit}} = 1$ in direction $\boldsymbol{\nu} = (0, -1)^T$. The Young's modulus of the material is $E = 10^2$ and the ersatz material (the white meshed area in Figure 5-8) has Young's modulus of $E = 10^{-3}$. Both have a Poisson ratio of $\nu = 0.3$. Again we assess the accuracy of fixed mesh methods using the solution on a fitted mesh shown in Figure 5-8(b) as a reference solution.

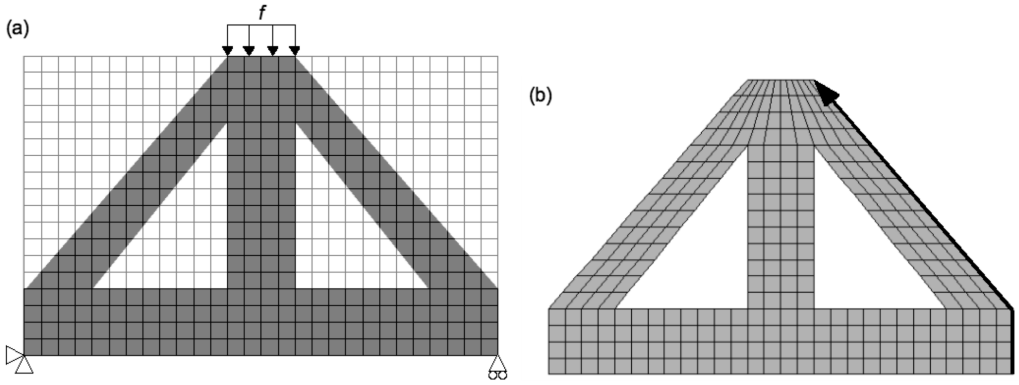
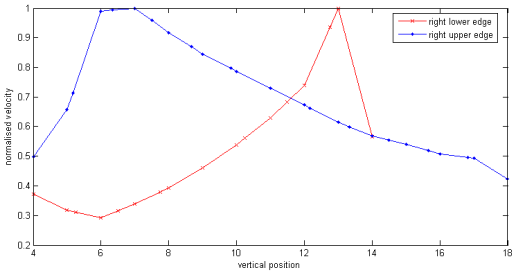


Figure 5-8: *The mesh and loading layout for the bridge problem.*

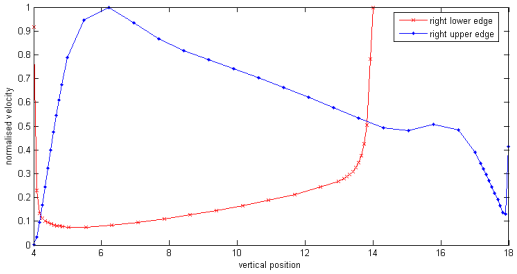
The bridge problem presents a new difficulty with topology optimisation methods. The sharp corners of the truss are points where strain concentrates, known to be points where corner singularities occur. This means that if the mesh is not sufficiently fine in the area local to the singularity then the error from the singularity will dominate the error everywhere. This is particularly bad when trying to find the shape sensitivity along the edges of the truss near these corners as we obtain a very inaccurate result. The profile for the right hand lower and upper diagonal edges is shown in Figure 5-9(a) for the fitted mesh shown in Figure 5-8(b).

Therefore we also compute a more accurate reference solution by using a fitted mesh and refinement near the corners. Figure 5-9(b) shows the shape sensitivity profile along the right diagonal edges in this case. It shows the correct concentration of strain in the corners and thus will achieve a better topology optimisation by adding material to round off these corners. Again, the fitted mesh calculations were done in ANSYS using biliear quadrilateral (Q1) elements.

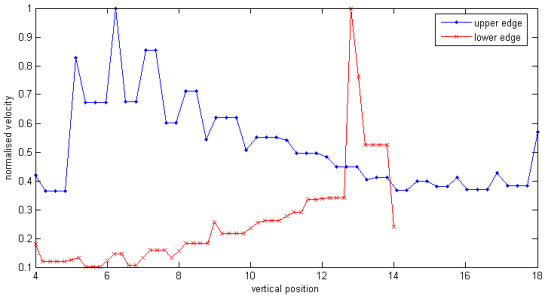
We compare our fixed mesh results to the profiles in Figure 5-9(a), 5-9(b). The results for the standard FEM with a 28×18 fixed uniform mesh are shown in Figure 5-9(c). We can see how the shape sensitivity again has a very poor accuracy compared to the refined fitted mesh data and has large jumps causing an oscillatory boundary to appear in the optimisation process. We remark that there is a dramatic improvement when



(a) Fitted mesh shape sensitivity profile.



(b) Fitted and refined mesh shape sensitivity profile.

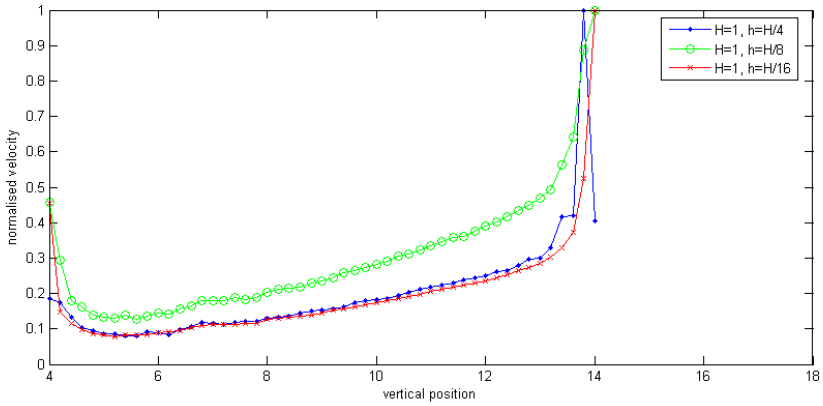


(c) Standard FEM uniform mesh shape sensitivity profile.

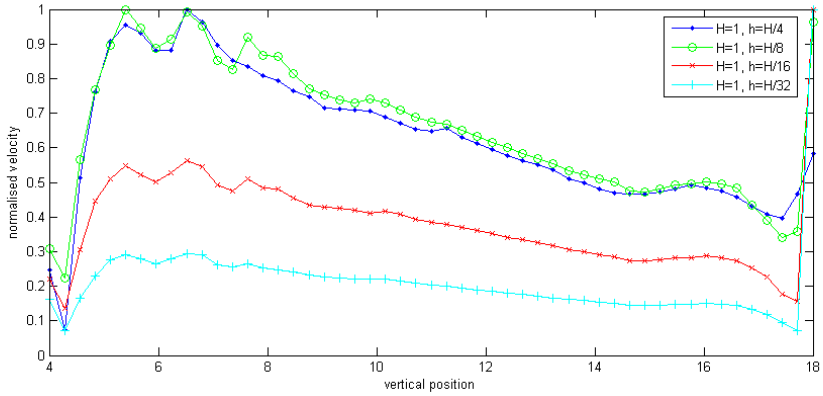
Figure 5-9: *The shape sensitivity along the right lower (red cross line) and upper (blue dot line) diagonal edges of the truss using a fitted mesh without and then with refinement at the corners.*

the adaptive MsFEM is used which we can see in Figure 5-10 when compared to Figure 5-9(b). The adaptive MsFEM is much smoother than the standard FEM and captures the correct behaviour at the singularities without any special action making it simpler to implement than extensive mesh refinement at the corners. We note however that the adaptive multiscale method does require a sufficiently fine subgrid diameter, h, to obtain an accurate solution to this problem at the end points of the edges. The shape sensitivity profile is shown for the lower edge in Figure 5-10(a) for decreasing subgrid h. We can see that for $h = H/4$ the adaptive method does not resolve the sensitivity at the ends very well, however this improves as h decreases to $H/16$.

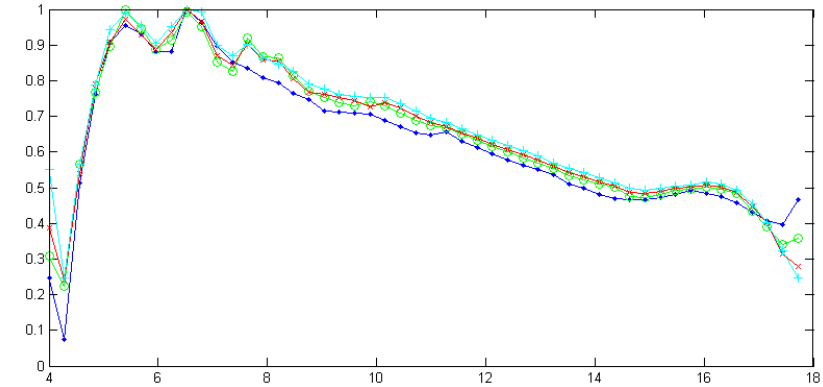
A difficulty occurs with the right upper diagonal edge, as the singularity is better resolved the solution becomes smoother as seen in Figure 5-10(b) but the stress concentration at the highest point distorts the normalised velocity. Figure 5-10(c) gives a more representative view where the last point has been removed and the shape sensitivity renormalised. The smoother solution is easier to see in this figure but we can see that the adaptive method is still having difficulty in the 6-8 vertical position region.



(a) AMsFEM uniform mesh shape sensitivity profile right lower diagonal edge.



(b) AMsFEM uniform mesh shape sensitivity profile right upper diagonal edge.



(c) AMsFEM uniform mesh shape sensitivity profile right upper diagonal edge with singularities removed.

Figure 5-10: *The shape sensitivity along the right lower (top) and upper (middle) diagonal edges of the truss using a uniform mesh with decreasing subgrid size h. The upper edge is recalculated without the corner points (bottom).*

The problem occurs because the singularities are close to the boundary, this problem can be resolved in some situations by extending the design domain by ghost material (or other appropriate extension) purely for the linear elasticity problem. We will explore this in the following benchmark example.

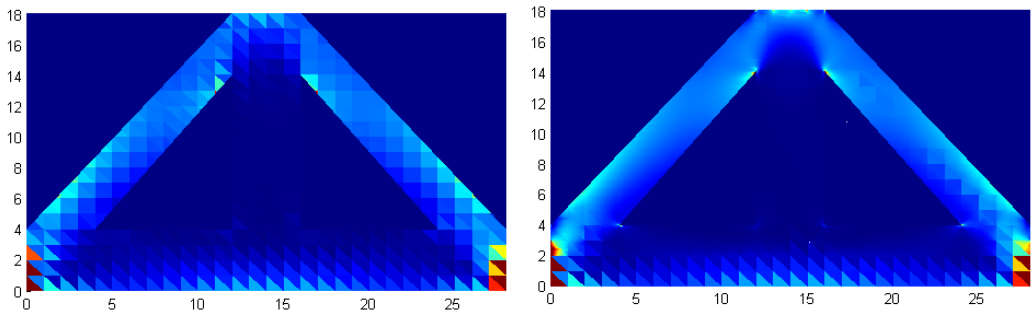


Figure 5-11: *Shape sensitivity comparison for the standard FEM (top) with $H = 1$ and the Adaptive MsFEM (bottom) with $H = 1$ and $h = H/16$.*

5.5.3 Membrane problem

The final benchmark example that we explore is a membrane structure under a shearing load. The design space is $\Omega = [0, 12] \times [0, 15]$ and the structure is given by $\Omega_S = \{x \in \Omega \,|\, 11x_1/12 \leq x_2 \leq x_1/3 + 11\}$. The membrane is fixed with $g = 0$ along $\Gamma_D = \{x \in \partial\Omega \,|\, x_1 = 0 \text{ and } 0 \leq x_2 \leq 11\}$. It is subject to the shearing force $f_{\text{unit}} = (0, 1)^T$ along $\Gamma_N = \{x \in \partial\Omega \,|\, x_1 = 12 \text{ and } 11 \leq x_2 \leq 15\}$ and traction free boundary conditions ($f_{\text{unit}} = (0, 0)^T$) on the remaining boundary. The material coefficients are as in the hole in a plate example (see Section 5.5.1).

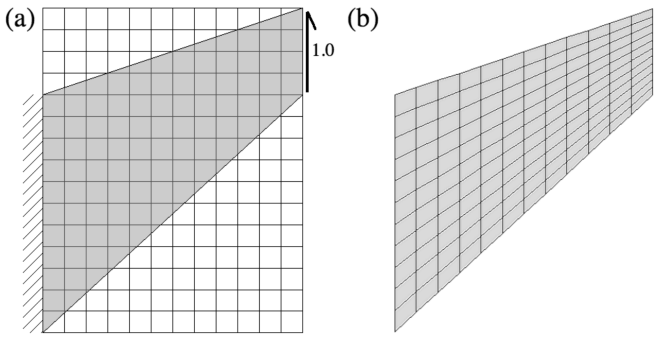
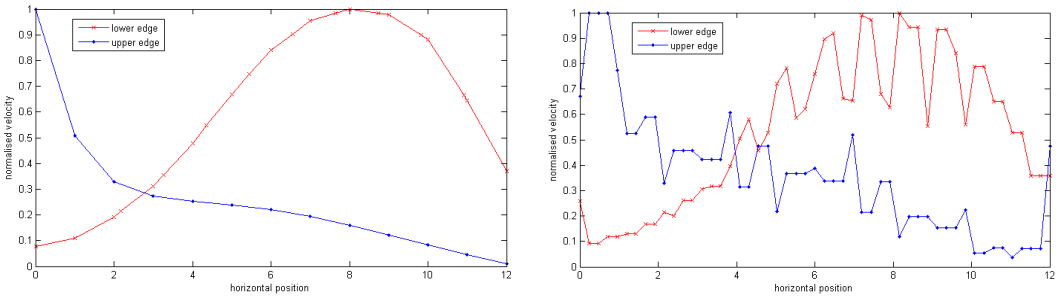


Figure 5-12: *The mesh and loading layout for the membrane problem.*

The shape sensitivities for the fitted mesh in Figure 5-12(b) are shown in Figure 5-13(a). Note that this solution has not had any refinement near the corner points but

still shows the correct behaviour as with a concentration of strain at the point $(0, 11)^T$. We can see in Figure 5-13(b) that the standard FEM for the fixed uniform 12×15 mesh follows a similar curve but has the usual oscillations in the shape sensitivity graph.



(a) ANSYS fitted mesh shape sensitivity profiles. (b) Standard FEM uniform mesh shape sensitivity profiles.

Figure 5-13: *The shape sensitivity along the lower and upper edges of the membrane using a fitted mesh in ANSYS (top) and a fixed uniform mesh with the standard FEM (bottom).*

In order to show what was meant by extending the design domain in the previous section, we consider extending Ω vertically by two elements for the Adaptive MsFEM. Thus we take $\widetilde{\Omega} = [0, 12] \times [-2, 17]$ and fill the extra domain with the ersatz material. This allows a clearer resolution of the corner singularities. The results for fixed H and decreasing subgrid h are shown in Figure 5-14 for the upper edge and Figure 5-15 for the lower edge. As with the previous benchmark examples we can see a much smoother sensitivity profile as well as a clearer resolution of the corner singularities.

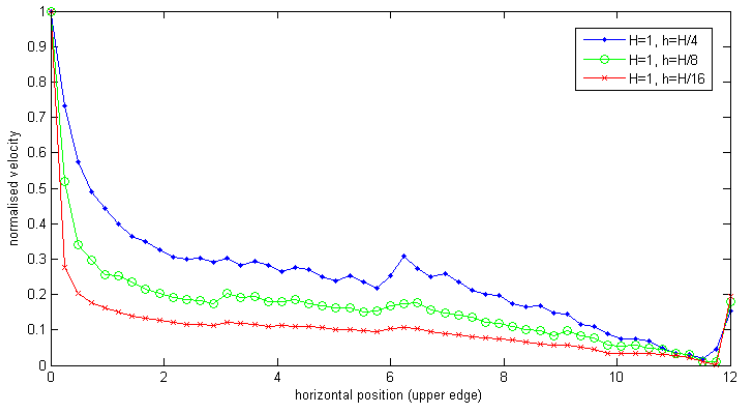


Figure 5-14: *The shape sensitivity along the upper edge of the membrane using AMsFEM and a uniform mesh with decreasing subgrid size h.*

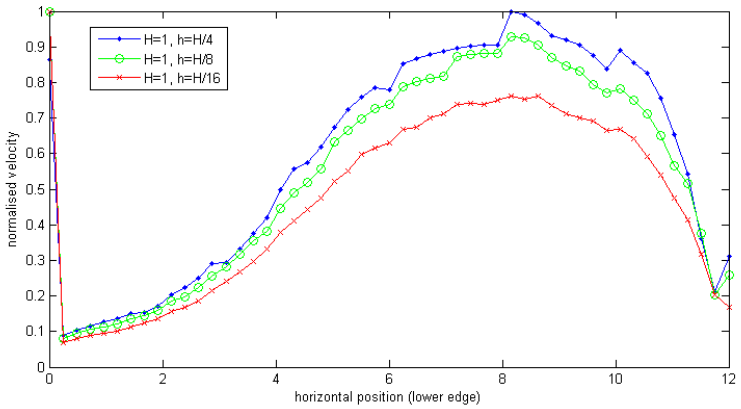


Figure 5-15: *The shape sensitivity along the lower edge of the membrane using AMsFEM and a uniform mesh with decreasing subgrid size h.*

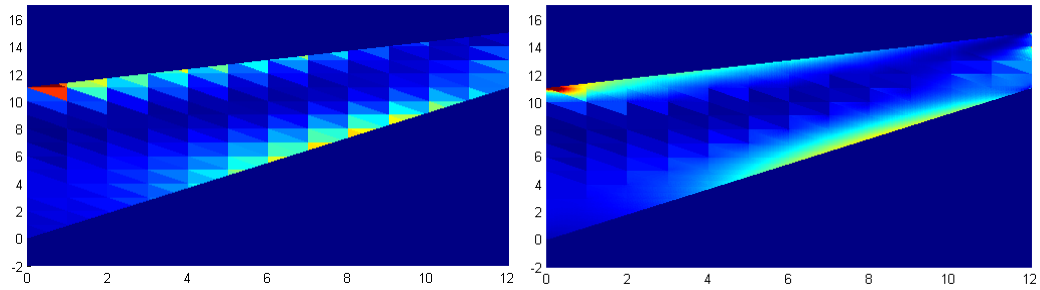


Figure 5-16: *Shape sensitivity comparison for the standard FEM (top) with $H = 1$ and the Adaptive MsFEM (bottom) with $H = 1$ and $h = H/16$ for the membrane problem.*

5.6 Summary

In this chapter we have shown a novel application of the adaptive multiscale finite element method to the linear elasticity problem and subsequent benefit to the structural optimisation process. The algorithm is easily generalisable to other bounded and coercive bilinear forms as we have shown in Section 5.2.

We showed how the adaptive multiscale finite element method provides a more accurate finite element approximation near the interface and thus for structural optimisation gives a smoother shape sensitivity profile along the interfaces without additional smoothing procedures. This was demonstrated by three benchmark examples in Section 5.5.

This chapter provides some initial work towards applying multiscale methods developed in porous media flow to other fields of engineering. Further research is required to explore the benefits of multiscale methods to structural optimisation.

6 Parallelisation of AMSFEM

6.1 Introduction

In our final chapter we consider the practical implementation of the adaptive multiscale finite element method. The process of solving local problems on each coarse grid element to find multiscale basis functions involves a lot of work that is not required by the standard finite element method. The adaptive method does however have the significant advantage that the error is smaller and the local problems can be solved in parallel. We will describe the parallel version of the AMsFEM algorithm indicating the specific points required to implement the communication between computer nodes. We explore some example results for the parallel code to demonstrate its scaling capability and we will finish the chapter by considering some possible enhancements for the AMsFEM algorithm.

6.2 The parallel adaptive multiscale finite element method

The main aim of the adaptive MsFEM is not to directly compete with the standard FEM in terms of execution time but rather to improve on the weaknesses of the standard method. The adaptive method takes significantly more time to complete when performed in serial due to the large number of local problems that need to be solved (even if the coarse mesh has a larger mesh diameter than for the standard FEM). Instead the adaptive method should be applied in situations where the standard FEM has a very poor convergence rate and thus requires an extremely fine mesh to obtain a required error, for example the interface problems with singularities or boundary layers (see Sections 4.6.3 and 4.6.4). As the number of degrees of freedom required to achieve a certain error level is much smaller for the AMsFEM compared to the standard FEM then the additional cost is offset by only considering a smaller problem.

This also introduces another situation when the AMsFEM is useful. When the size of the stiffness matrix that arises from a fine mesh becomes too large for current computational resources then the AMsFEM can be used. If the standard FEM converges at a rate slower than the AMsFEM (e.g. for interface problems where the standard FEM is only $O(H)$ in the L_2 norm compared to $O(H^2)$ for the AMsFEM) then to achieve the same error we can use a coarse mesh and thus have a smaller stiffness matrix. This may be small enough to fit into the available memory resources. Note that the AMsFEM will require more out-of-core resources than the standard FEM in order to store the basis functions but it would still offer an advantage for large problems.

The other possibility is to use the AMsFEM when the problem has to be solved repeatedly. For example as part of a porous media flow problem where the same field is used for many time steps, or in a linear elasticity problem when multiple loading conditions or multiple boundary conditions are examined. There is also an advantage in each of these situations if the coefficient field changes slightly or only in certain regions as you can retain some basis functions or use them as a starting point for the iterative process thereby requiring fewer iterations, both of which save computing time.

The final instance where the AMsFEM is useful is in interface problems where the inclusions are relatively large compared to the size of the coarse mesh. In this situation then the number of cut elements that require a local problem to be solved is $O(H^{-1})$. Depending on the expense of the local solve itself, the cost of the local problems should be comparable to the cost of the solution of the global problem (at least $O(H^{-2})$ even with an optimal iterative solver).

As well as the advantage of the AMsFEM in these situations, it is also possible to improve its performance by implementing a parallel version of the algorithm. The parallel AMsFEM (ParAMsFEM) follows the same algorithm as outlined in Algorithm 2 but distributes the local problems across many processors. A flowchart of the algorithm is shown in Figure 6-1 indicating which parts of the algorithm can be done in parallel and which parts form the bottlenecks and can only be done in serial or require communication.

Figure 6-1 also helps to show the difference between the original ALG-MsFE methods by Durlofsky, Efendiev and Ginting [36] and the enhanced version given by AMsFEM. The ALG-MsFE methods use only the first stage of parallelisation where the local problems (Algorithm 1, Section 4.3.4) are solved to get non-conforming $\Phi^{\mathrm{MS}}_{\tau,i}$. We enhance the method by introducing the edge averaging of basis functions and then the second stage of parallelism with the local problem being resolved on each element. This helps to

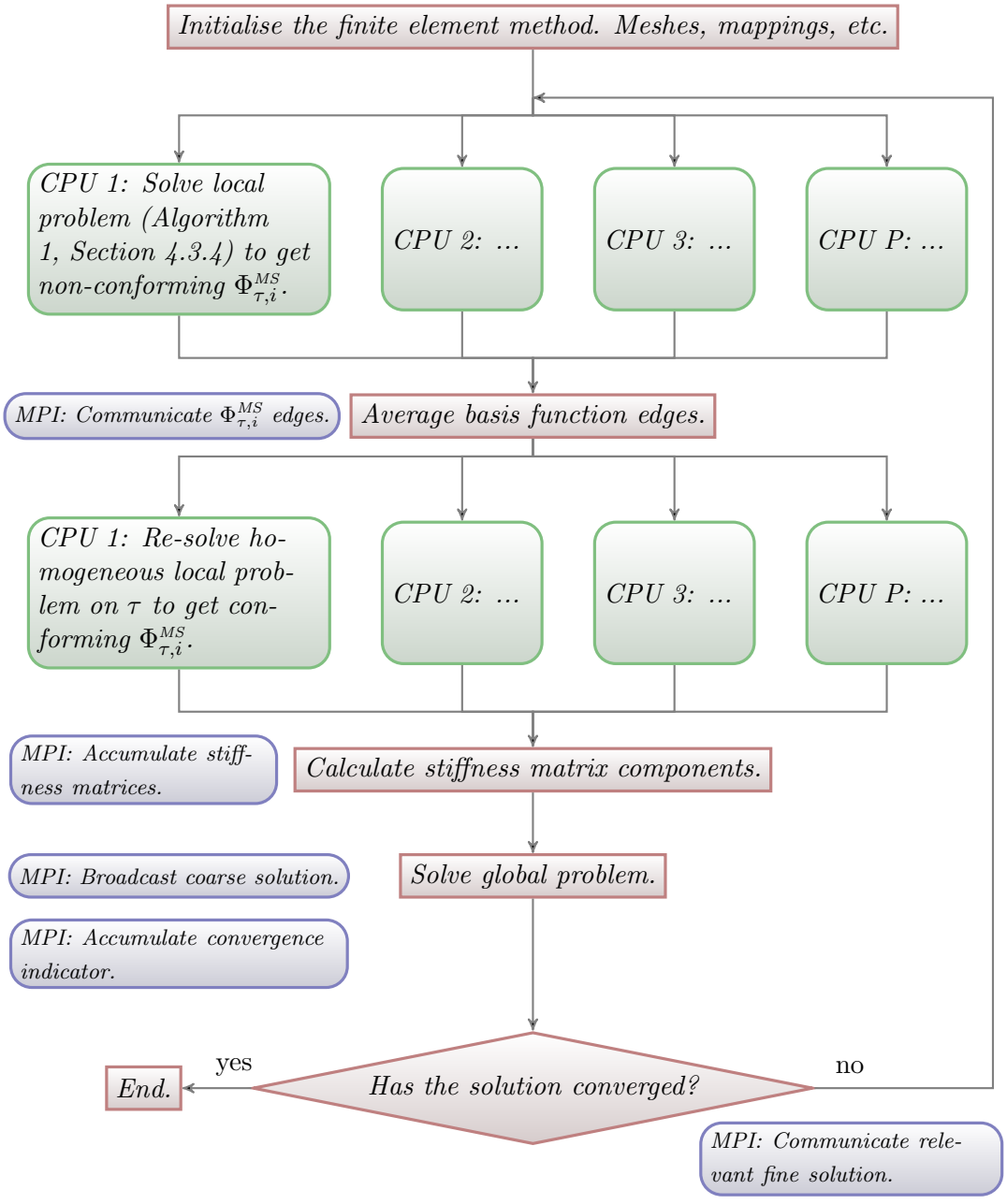


Figure 6-1: *Flowchart for the parallel adaptive multiscale finite element method indicating which parts of the algorithm can be parallelised and which are serial only.*

reduce the number of iterations required and experimentally gives more stability to the method. The drawback is that the edge averaging process introduces a new bottleneck to the algorithm as basis functions on different processors have to be exchanged.

Although Figure 6-1 gives a good reference as to how the AMsFEM algorithm is parallelised, we consider some of the details below regarding the bottlenecks and communication:

- **Partitioning the local problems.** The first major consideration for the parallel model is how the local problems are distributed to different processors. In our implementation this is done crudely by having one master node which handles standard elements and solving the global problem while the remaining multiscale elements are numbered linearly and shared equally between a set of worker nodes. This master-worker model has the disadvantages that the master is idle for a lot of the time and the cut elements are not distributed in the best fashion between workers. A partitioning algorithm could ensure that there is as little communication overlap between processors as possible thus improving performance. The best way to do this would be to utilise some partitioning software such as METIS [54] and in fact the partitioning can be done in parallel using ParMETIS, thus reducing the percentage of serial code and increasing the maximum speed up according to Amdahl's law.

- **Averaging the basis function edges.** The second bottleneck concerns averaging the basis function edges to produce a conforming finite element method. In the algorithm outlined in Figure 6-1 all the local solves are completed and then the values along the element edges are shared between processors. This is not optimal as we have to wait for all communication to finish before proceeding. The problem is alleviated by using asynchronous communication. Since each element has a finite number of neighbours then we can determine the size of a communication buffer a priori based on the partitioning in the previous bullet point. The asynchronous sends and receives are initiated and allowed to fill the buffer with the basis function values in any element order. The algorithm must then wait for all of these communications to finish before proceeding.

- **Building the stiffness matrices and solving the global problem.** The ParAMsFEM algorithm we have described uses a serial matrix solver for the global problem. As such the global stiffness matrix must be assembled on the master node. This is done by sending each local stiffness matrix back to the head node, again in an asynchronous fashion, and performing the assembly on the head

node. This has the drawback of a large overhead for initiating communications for each cut element. Another possibility would be to assemble a stiffness matrix on each processor and accumulate it on the head node. The drawback then is that the size of the sparse matrix created for each processor is unknown in advance unless a worst case scenario is used. Our implementation used the Intel MKL library and the PARDISO direct solver. These could be replaced by other solvers such as the HSL (Harwell Subroutine Library) routines or in fact a parallel matrix system solver such as MUMPS. A parallel solver would also reduce communication if only parts of the matrix had to be stored on each processor. It would also remove the largest part of serial code slowing the algorithm down, as we will see in the following section on numerical results.

- **Distributing the coarse solution and accumulating the stopping condition.** After solving the global problem on the coarse mesh then the coarse solution has to be distributed to all the other processors in order to calculate a stopping condition. In our implementation this was taken to be the relative change in the L_2 norm of the fine scale solution. The coarse solution is broadcast to each processor from the head node. If a parallel solver were used for the global problem then a more complicated set up would be required using asynchronous transfers as in the basis function averaging stage. The error indicator is then also accumulated by each processor sending its indicator value to the head node where it is reduced to a single value.

- **Distributing the fine scale solution.** The final bottleneck to the algorithm is the distribution of the fine solution along the edges of elements that are required by other processors when solving the local problems on an extended domain. If the stopping condition is not satisfied then the head node broadcasts an instruction to get each process to iterate again, however, in order to use the projection $\mathcal{P}_{i,\tilde{\tau}}$ on the extended element $\tilde{\tau}$ we first need the values of u_H^{MS} along $\partial\tilde{\tau}$. Again a good partitioning by METIS should reduce this cost.

6.3 Numerical results

To give an indication of the scaling capability of the parallel adaptive MsFEM we consider two numerical experiments. We explore timings for the random field zero source mixed boundary condition problem at the end of Section 4.6.5 with length scale $\lambda = 0.01$ and standard deviation $\sigma = 2.5$. We consider two different mesh sizes, $H = 1/32$ and $H = 1/64$ and record the execution times. Using these times (T_P)

and corresponding number of processors (P), we can calculate the speed up (S_P) and efficiency (E_P) by

$$S_P = \frac{T_1}{T_P} \qquad\qquad E_P = \frac{S_P}{P} \tag{6.1}$$

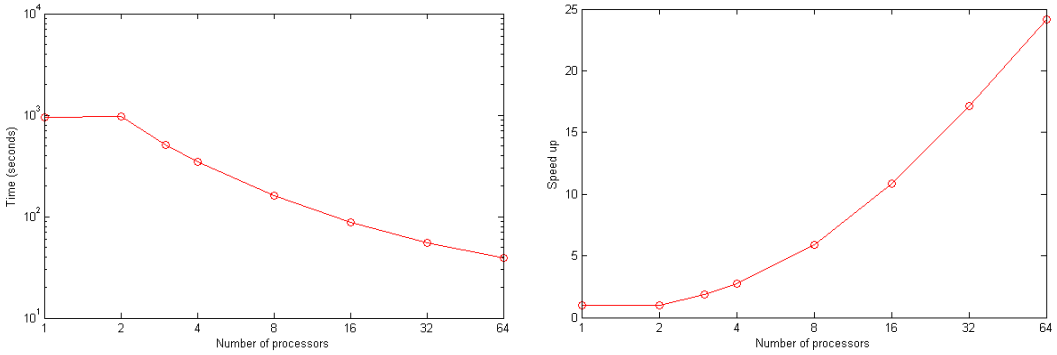
Our first set of results is shown in Table 6.1 for the case when $H = 1/32$. The subgrid mesh size was taken to be $h = H/8$ and the iterative process required 5 iterations. The errors for this experiment (using 8 processors) are given in Table 4.22. For $H = 1/32$ the AMsFEM produces an error of 4.8243×10^{-3} which is compared to the standard FEM error of 4.2403×10^{-3} for $H = 1/256$. The standard FEM takes 34 seconds to complete while the AMsFEM takes significantly longer in serial. However, as we use more processors for the local problems then the cost becomes comparable with the AMsFEM reducing to 39 seconds. We note that both algorithms solve the global problem in serial, a fairer comparison would be to examine versions of each algorithm that also utilise a parallel global solve. This however is not the key issue here as the adaptive method would be better used when repeated experiments have to be performed that can utilise the same basis functions. We simply show that parallelisation offers the capability to perform the adaptive method in a comparable time to current serial technology. The speedup is greater for problems that have a slower rate of convergence of the error with the standard FEM compared to the improved rates of the AMsFEM.

Num CPUs, P	Time, T_P (secs)	Speed up, S_P	Efficiency, E_P (%)
1	943	1.0000	100.0000
2	967	0.9752	48.7590
3	507	1.8600	61.9987
4	348	2.7098	67.7443
8	160	5.8937	73.6719
16	87	10.8391	67.7443
32	55	17.1455	53.5795
64	39	24.1795	37.7804

Table 6.1: *Timings and statistics for the parallel adaptive MsFEM for the mixed boundary condition random field problem in Section 4.6.5 for $H = 1/32$ and $h = H/8$.*

The results in Table 6.1 are shown graphically in Figure 6-2. In Figure 6-2(a) we can see the clear effect of having a master and workers scheme for the AMsFEM algorithm. The cost on 2 processors is approximately the same as the cost on one node since the master node is idle whilst the worker solves the local problems. We can see that there is not a large amount of overhead associated with communication between the workers and head node in this instance. What we do see is that the time does not decrease linearly although it is near linear up to 16 processors. The reduction in speedup is because the

global coarse grid problems are still solved in serial and thus this becomes the dominant cost of the AMsFEM. Larger experiments are required to show the plateau expected under Amdahl's law as the serial part of the algorithm becomes the limiting factor. We will see this plateau in our next timing experiment.



(a) Timing results.

(b) Speed up results.

Figure 6-2: *Graphical plot of the results in Table 6.1 showing timings and speed up.*

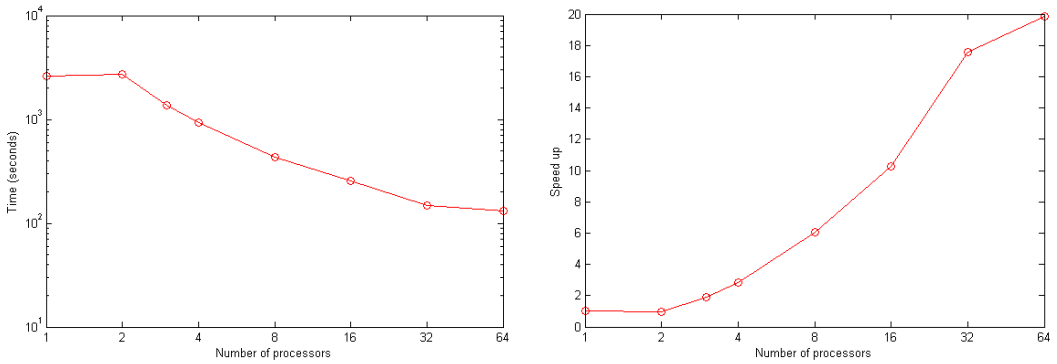
For our second experiment we consider the same problem as before but now with $H = 1/64$. This is for the purpose of showing that the AMsFEM competes more strongly with the standard FEM for larger problems. Even if we suppose that the standard FEM was converging at an $O(H^2)$ rate between $H = 1/256$ and $H = 1/512$ (this being the best case when the mesh resolves the coefficient but normally would only be $O(H)$) then we would have an error of the order of 10^{-3} which would be comparable to the AMsFEM with $H = 1/64$. The standard FEM required 298 seconds to complete for $H = 1/512$ which, with sufficient parallelisation of the local problems, we can achieve a better time. Note again that both methods use a serial global solve.

Num CPUs, P	Time, T_P (secs)	Speed up, S_P	Efficiency, E_P (%)
1	2619	1.0000	100.000
2	2707	0.9675	48.375
3	1360	1.9257	64.191
4	933	2.8071	70.177
8	433	6.0485	75.606
16	255	10.2706	64.191
32	149	17.5772	54.929
64	132	19.8409	31.001

Table 6.2: *Timings and statistics for the parallel adaptive MsFEM for the mixed boundary condition random field problem in Section 4.6.5 for $H = 1/64$ and $h = H/8$.*

As in the first experiment we display the results in Table 6.2 graphically in Figure 6-3. The results show much the same behaviour as before but we see a much quicker drop

off in speed up towards the higher processor numbers. This shows that as the problem grows in size then the global problems are becoming a much more significant part of the algorithm and thus should also be parallelised. We note that the drop in performance is also due to the fact that the number of elements is fixed and therefore ultimately it can only be split between a finite number of processors. We also see that the gain from splitting the local problems between processors is also starting to be outweighed by the cost of communication between processors. Hiding the communication time behind the processing time could alleviate this problem and provide a possible enhancement to the algorithm.



(a) Timing results.

(b) Speed up results.

Figure 6-3: *Graphical plot of the results in Table 6.2 showing timings and speed up.*

6.4 Adaptive multiscale method algorithm enhancements

The parallel adaptive multiscale method as stated in Section 6.2 is a general description of the method and allows some flexibility. There are several options for improving parallel performance as well as increasing the order of convergence.

The first option to improve parallel performance is to incorporate the bottleneck points into the parallel implementation. The parallel algorithm given in Figure 6-1 seeks to solve all of the local problems using Algorithm 1 in parallel first, then exchange basis function values along element edges before averaging the basis functions along neighbouring edges. Each processing node must wait for every other node to complete its communications before averaging edges. This could be improved by employing asynchronous communications.

The parallel algorithm could be modified to solve local problems on elements that are involved in communication between processing nodes first, these communications could

be started and the remaining local problems solved whilst waiting for the communication to complete. Similarly, conforming basis functions can be calculated on elements that do not require communication first, and then the remaining basis functions left until the end after communication is complete. This in effect hides the communication time behind processing time and would give a significant advantage when the amount of communication overlap is small in comparison to the total number of basis functions to be calculated.

The biggest bottleneck in the algorithm is that the global problem must be solved for every iteration. This requires the local element stiffness matrices from each processing node to be assembled, resulting in a large amount of communication back to a head node that calculates the solution of the resulting matrix system. The matrix solve can be improved by employing a parallel solver, thus making use of all the other nodes. If a parallel solver were employed for the global problem then partitioning elements for the local solves in such a way that little communication of local element stiffness matrices was needed would also improve performance.

As we have seen in the numerical results in Section 4.6 and 5.5, the adaptive multiscale finite element method is better than the standard FEM but does not always achieve an optimal rate of convergence. This is due to the possibility that the local subgrid may be inaccurate. In our implementation, the local solves were done using the Immersed FEM [62] for the interface problems and standard FEM for the random field problems. Since any suitable method can be used to solve the local problems, it is possible to envisage a multilevel adaptive MsFEM where the local problems are again solved by the adaptive MsFEM down to a sufficiently fine level where optimal convergence is restored.

An easier method to improve the accuracy of the local solves for interface problems is to employ conventional h-refinement where a rough mesh is applied to the element and it is refined near the interface. Also r-refinement could be used to move the mesh nodes around so that the interface is better resolved. The final possibility is to introduce p-refinement as well, all of our experiments simply used linear subgrid finite elements but this could be upgraded to higher polynomials. These ideas on refinement could also apply in more general cases such as the random field problems by employing an error indicator and refining based on the indicator. Note that all of this refinement could be done in parallel as it is local to the specific coarse mesh element. Each coarse mesh element could also use different refinement processes provided that the values of the basis functions along the edges are continuous and can be exchanged.

The last option for enhancement is to use an indicator to control the number of iter-

ations of basis functions on each element seperately. Therefore if the indicator does not change significantly between one iteration and the next, for example the $H^1(\Omega), \alpha$ energy norm, then the iterative process could stop on that particular element. This will reduce the computing time as the adaptive MsFEM requires fewer iterations when the coarse mesh is of a similar scale to the permeability field $\mathcal{A}$, thus if $\mathcal{A}$ contains regions of different scales then different basis functions will require varying numbers of iterations. This enhancement has recently been implemented by Hajibeygi and Jenny in [43] to create an adaptive iterative multiscale finite volume method that builds on the work in [53] for the original multiscale finite volume method.

6.5 Summary

In our last chapter we have shown how, even though the serial AMsFEM algorithm is more computationally intensive than the standard finite element method for a comparable error in all but a few situations, the AMsFEM algorithm is ideally suited to parallelisation. While the parallel AMsFEM algorithm is not embarrisingly parallel it is at least a coarse-grained parallel algorithm. In Section 6.3 we provided scaling results for an example problem and showed that the parallel version provides a significant advantage for reducing the execution time over the serial algorithm. Further enhancements would improve the parallel algorithm and reduce the execution time further.

7

Conclusions and further work

In this thesis we have successfully shown that, under certain assumptions, the finite element error for the high contrast elliptic interface problem is independent of the contrast in the coefficient that the bilinear form depends on. Thus is not as pessimistic as previous results suggest. In Chapter 2 we gave a full description of the interface problem and showed how, through the use of Galerkin optimality, it was possible to bound the finite element error by bounding the error between the true solution and a construction. The key to the constructed function is to approximate the gradient of the solution well in regions where the coefficient is largest in order to prove a finite element error that is independent of the contrast, this was shown in Section 2.3. This was extended by a technical proof in Section 2.4 to the whole domain. Numerical results to suggest that the bounds in Chapter 2 are sharp are given in Section 4.6 where the standard finite element method is tested against the more advanced multiscale finite element method.

In Chapter 3 we reviewed the work of Chu, Graham and Hou in [27] with regard to their multiscale finite element method. We gave a clearer insight into their work by describing their method and reviewing their analysis, giving generalisations where possible (see Lemma 3.19) with the view that future work may extend the analysis of a priori contrast independent local boundary conditions further. The key to all of the a priori error estimates is the introduction of a coefficient explicit regularity theory in the appendix of [27], the proof however was only for a single inclusion and so in this thesis we have extended it to multiple inclusions. While we used the regularity theory for the standard finite element error estimates in Chapter 2, it also led to the creation of a relative error estimate for the finite element error. One of the assumptions regarding the coefficient $\mathcal{A}$ in Chapters 2, 3 and [27] was that $\mathcal{A} \geq 1$, meaning that while the estimates are contrast independent they are not coefficient independent. The relative error bound explicitly shows that the finite element error is independent of the coefficient because the solution blows up as $\min_{x\in\Omega} \mathcal{A}(x) \to 0$, corresponding to a loss

of ellipticity.

While the algorithm for finding these artificial local boundary conditions is simple, the analysis is extremely complicated and limited by the geometrical assumptions. In Chapter 4 we instead looked at generalising the local-global approach by Durlofsky, Efendiev and Ginting in [36]. Their approach was to find the artificial local boundary conditions iteratively in order to obtain multiscale basis functions. In Section 4.5 we gave a more general framework that encompasses their methods but with the drawback that their best method is non-conforming, we showed in Section 4.5.4 that their non-conforming method could be enhanced to produce a conforming method that still retained the superior convergence of their method. Again the enhanced method was shown to fit within the general framework algorithm. The real benefit of this adaptive approach is that it can be applied to any high contrast elliptic problem and is not limited to only interface problems. We showed in Section 4.6 that the method performs extremely well numerically over the standard finite element method and is also contrast independent, especially in situations where the coefficient contains a corner singularity or boundary layer.

The novel application of the research in this thesis was described in Chapter 5 where the multiscale finite element methods, typically applied to porous media flow problems, were applied to the linear elasticity problem for the structural optimisation process. Conventionally the idea of using multiscale basis functions has been applied to high contrast heterogeneous elliptic problems representing the permeability field of a rock structure, we instead applied the same techniques gained from generalising the description of the local-global methods in Chapter 4 to the field of mechanical engineering. The structural optimisation process using the level-set method presented a perfect example of a high-contrast elliptic interface problem as outlined in Chapter 2. We showed, via three benchmark examples, that the sensitivity profile along the boundary of the structure is much more accurate using the adaptive multiscale FEM than the standard FEM and removes the need for heuristic smoothing techniques widely used within the engineering community.

While the adaptive multiscale finite element method is more accurate it is also much more computationally expensive. For this reason it was necessary to construct a parallel version of the algorithm where the calculation of the local multiscale basis functions can be distributed to many computers. Note that the real advantage of the multiscale finite element method is when the multiscale basis functions can be repeatedly used, therefore the expensive step only occurs once and then the adaptive method out performs the standard FEM significantly. In Chapter 6 we described how the adaptive method could

be parallelised, showed via a scaling experiment that the parallelisation is good and also suggested several methods for improving the performance.

At the beginning of the thesis we set out to prove contrast independent finite element error estimates. We have explored multiscale finite element methods and developed a new adaptive approach to define a conforming multiscale finite element method, showing numerically the superior convergence of the adaptive method over the standard finite element method. We have showed a novel application of the multiscale method to linear elasticity problems arising in structural engineering and we have shown how to implement the method on a parallel computer. Although this thesis has successfully managed to further our knowledge in these areas it also presents many new questions for future research.

The first area is to extend the a priori finite element error estimates in Chapter 2 to 3D, this will require a different technique as the proof in this thesis uses the Sobolev embedding theorem from $H^{1+\epsilon}$ to L_∞ in $\mathbb{R}^2$ but in $\mathbb{R}^3$ $H^{\frac{3}{2}+\epsilon}$ embeds into L_∞ however the solution is still only in $H^{\frac{3}{2}-\epsilon}$. The second area to extend in is by finding more apriori contrast independent local boundary conditions as in Chapter 3 in order to handle other configurations of how the interface cuts the element. A greater use of research effort would be to analyse the convergence properties of the Adaptive Multiscale FEM in Chapter 4 which is more widely applicable to any high contrast elliptic problem. One strategy for this is to prove that the true solution forms a fixed point of the method and then prove the method is a nonlinear contraction mapping. The difficulty with this idea is to show the influence of the projection in (4.5) on the finite element error which is as yet unknown. The other possible direction for research would be to develop a 3D version of the adaptive multiscale method, for this a two dimensional version of the projection function on the faces of elements needs to be developed. Other more immediate areas for research involve incorporating the adaptive method into the whole structural optimisation process and study how it affects the number of optimisation steps required. The final area for further work is to implement the extensions to the parallel adaptive method outlined in Section 6.4.

Bibliography

[1] J. Aarnes. On the Use of a Mixed Multiscale Finite Element Method for Greater-Flexibility and Increased Speed or Improved Accuracy in Reservoir Simulation. *Multiscale Modeling & Simulation*, 2:421, 2004.

[2] J. Aarnes and Y. Efendiev. An adaptive multiscale method for simulation of fluid flow in heterogeneous porous media. *Multiscale Modeling and Simulation*, 5(3):918, 2006.

[3] J. Aarnes, V. Kippe, and K. Lie. Mixed multiscale finite elements and streamline methods for reservoir simulation of large geomodels. *Advances in Water Resources*, 28(3):257–271, 2005.

[4] J. Aarnes, S. Krogstad, and K. Lie. A hierarchical multiscale method for two-phase flow based upon mixed finite elements and nonuniform coarse grids. *Multiscale Modeling and Simulation*, 5(2):337–363, 2007.

[5] L. Adams and Z. Li. The immersed interface/multigrid methods for interface problems. *SIAM Journal on Scientific Computing*, 24(2):463–479, 2003.

[6] G. Allaire. Homogenization and two-scale convergence. *SIAM Journal on Mathematical Analysis*, 23:1482, 1992.

[7] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve. Shape optimization by the homogenization method. *Numerische Mathematik*, 76(1):27–68, 1997.

[8] G. Allaire and M. Briane. Multiscale convergence and reiterated homogenisation. *Royal Society(Edinburgh), Proceedings, Section A*, 126:297–342, 1996.

[9] G. Allaire, F. De Gournay, F. Jouve, and A. Toader. Structural optimization using topological and shape sensitivity via a level set method. *Control and Cybernetics*, 34(1):59, 2005.

[10] G. Allaire, F. Jouve, and A. Toader. Structural optimization using sensitivity analysis and a level-set method. *Journal of Computational Physics*, 194(1):363–393, 2004.

[11] I. Babuška. The finite element method for elliptic equations with discontinuous coefficients. *Computing*, 5(3):207–213, 1970.

[12] I. Babuška. Solution of interface problems by homogenization. i. *SIAM Journal on Mathematical Analysis*, 7:603, 1976.

[13] I. Babuška. Solution of interface problems by homogenization. ii. *SIAM Journal on Mathematical Analysis*, 7:635, 1976.

[14] I. Babuška. Solution of interface problems by homogenization. iii. *SIAM Journal on Mathematical Analysis*, 8:923, 1977.

[15] I. Babuška, U. Banerjee, and J. Osborn. Generalized finite element methods-main ideas, results and perspective. *Int. J. Comput. Methods*, 1(1):1–37, 2004.

[16] I. Babuška, G. Caloz, and J. Osborn. Special finite element methods for a class of second order elliptic problems with rough coefficients. *SIAM Journal on Numerical Analysis*, 31(4):945–981, 1994.

[17] I. Babuška and J. Osborn. Generalized finite element methods: their performance and their relation to mixed methods. *SIAM Journal on Numerical Analysis*, 20(3):510–536, 1983.

[18] C. Bernardi and R. Verfürth. Adaptive finite element methods for elliptic equations with non-smooth coefficients. *Numerische Mathematik*, 85(4):579–608, 2000.

[19] J. Bourgat. Numerical experiments of the homogenization method. *Computing Methods in Applied Sciences and Engineering, 1977, I*, pages 330–356, 1979.

[20] S. Brenner and L. Scott. *The mathematical theory of finite element methods*. Springer Verlag, 2008.

[21] F. Brezzi. Recent results in the treatment of subgrid scales. In *ESAIM: Proceedings*, volume 11, pages 61–84. edpsciences. org, 2002.

[22] F. Brezzi, T. Hughes, L. Marini, A. Russo, and E. Süli. A priori error analysis of residual-free bubbles for advection-diffusion problems. *SIAM Journal on Numerical Analysis*, 36(6):1933–1948, 1999.

[23] Y. Chen and L. Durlofsky. Adaptive local–global upscaling for general flow scenarios in heterogeneous formations. *Transport in Porous Media*, 62(2):157–185, 2006.

[24] Y. Chen, L. Durlofsky, M. Gerritsen, and X. Wen. A coupled local-global upscaling approach for simulating flow in highly heterogeneous formations. *Advances in Water Resources*, 26(10):1041–1060, 2003.

[25] Z. Chen and J. Zou. Finite element methods and their convergence for elliptic and parabolic interface problems. *Numerische Mathematik*, 79(2):175–202, 1998.

[26] I. Chern et al. A coupling interface method for elliptic interface problems. *Journal of Computational Physics*, 225(2):2138–2174, 2007.

[27] C. Chu, I. Graham, and T. Hou. A new multiscale finite element method for high-contrast elliptic interface problems. *Math. Comp*, 79:1915–1955, 2010.

[28] J. Chu, Y. Efendiev, V. Ginting, and T. Hou. Flow based oversampling technique for multiscale finite element methods. *Advances in Water Resources*, 31(4):599–608, 2008.

[29] P. Ciarlet. *The finite element method for elliptic problems*. North-Holland, 1978.

[30] R. Cook. *Finite element modeling for stress analysis*. Wiley, 1995.

[31] R. Dautray, J. Lions, M. Artola, and M. Cessenat. *Mathematical analysis and numerical methods for science and technology: Physical origins and classical methods*. Springer Verlag, 1990.

[32] M. Dumett and J. Keener. An immersed interface method for solving anisotropic elliptic boundary value problems in three dimensions. *SIAM Journal on Scientific Computing*, 25:348, 2003.

[33] P. Dunning, H. Kim, and G. Mullineux. Error Analysis of Fixed Grid Formulation for Boundary Based Structural Optimisation. In *7th ASMO UK Conference on Engineering Design Optimization, Bath, UK*. University of Bath, July 2008.

[34] P. Dunning, H. Kim, and G. Mullineux. Two-Dimensional Fixed Grid Based Finite Element Structural Analysis. In *12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Victoria, British Columbia, Canada*. University of Bath, Sept 2008.

[35] T. Dupont and R. Scott. Polynomial approximation of functions in Sobolev spaces. *Mathematics of Computation*, 34(150):441–463, 1980.

[36] L. Durlofsky, Y. Efendiev, and V. Ginting. An adaptive local–global multiscale finite volume element method for two-phase flow simulations. *Advances in Water Resources*, 30(3):576–588, 2007.

[37] Y. Efendiev and T. Hou. *Multiscale finite element methods: theory and applications*. Springer Verlag, 2009.

[38] B. Engquist and O. Runborg. Wavelet-based numerical homogenization with applications. *Multiscale and Multiresolution Methods: Theory and Applications*, page 97, 2002.

[39] R. Fedkiw, T. Aslam, B. Merriman, and S. Osher. A non-oscillatory Eulerian approach to interfaces in multimaterial flows (the ghost fluid method). *Journal of Computational Physics*, 152(2):457–492, 1999.

[40] M. García-Ruíz and G. Steven. Fixed grid finite elements in elasticity problems. *Engineering Computations*, 16(2):145–164, 1999.

[41] G. Golub and C. Van Loan. *Matrix computations*, volume 3. Johns Hopkins Univ Pr, 1996.

[42] M. Graphics. http://www.mentor.com/company/news/flomerics-flothermpcb-thermal-software-is-finalist-in-iecs-2007-designvision. online. accessed 18th May 2011.

[43] H. Hajibeygi and P. Jenny. Adaptive iterative multiscale finite volume method. *Journal of Computational Physics*, 2010.

[44] Z. Hassan, N. Allec, L. Shang, R. Dick, V. Venkatraman, and R. Yang. Multiscale thermal analysis for nanometer-scale integrated circuits. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 28(6):860–873, 2009.

[45] P. Henning and M. Ohlberger. The heterogeneous multiscale finite element method for elliptic homogenization problems in perforated domains. *Numerische Mathematik*, 113(4):601–629, 2009.

[46] A. Heyden. http://www.cse.sc.edu/ heyden/Multi-ScaleModelling.html. online. accessed 18th May 2011.

[47] V. Hoang and C. Schwab. High-dimensional finite elements for elliptic problems with multiple scales. *Multiscale Modeling & Simulation*, 3:168, 2005.

[48] T. Hou, Z. Li, S. Osher, and H. Zhao. A hybrid method for moving interface problems with application to the Hele-Shaw flow. *Journal of Computational Physics*, 134(2):236–252, 1997.

[49] T. Hou and X. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *Journal of Computational Physics*, 134(1):169–189, 1997.

[50] T. Hou, X. Wu, and Z. Cai. Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients. *Mathematics of Computation*, 68(227):913–943, 1999.

[51] J. Huang and J. Zou. Some New A Priori Estimates for Second-Order Elliptic and Parabolic Interface Problems. *Journal of Differential Equations*, 184(2):570–586, 2002.

[52] T. Hughes, G. Feijóo, L. Mazzei, and J. Quincy. The variational multiscale method–a paradigm for computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 166(1-2):3–24, 1998.

[53] P. Jenny, S. Lee, and H. Tchelepi. Multi-scale finite-volume method for elliptic problems in subsurface flow simulation. *Journal of Computational Physics*, 187(1):47–67, 2003.

[54] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359, 1999.

[55] H. Kim, P. Clement, and J. Cunningham. Investigation of cancellous bone architecture using structural optimisation. *Journal of Biomechanics*, 41(3):629–635, 2008.

[56] H. Kim, M. Garcia, O. Querin, G. Steven, and Y. Xie. Introduction of fixed grid in evolutionary structural optimisation. *Engineering Computations*, 17(4):427–439, 2000.

[57] H. Kim, O. Querin, G. Steven, and Y. Xie. Improving efficiency of evolutionary structural optimization by implementing fixed grid mesh. *Structural and Multidisciplinary Optimization*, 24(6):441–448, 2002.

[58] R. Leveque and Z. Li. The immersed interface method for elliptic equations with discontinuous coefficients and singular sources. *SIAM Journal on Numerical Analysis*, 31(4):1019–1044, 1994.

[59] J. Li, J. Melenk, B. Wohlmuth, and J. Zou. Optimal a priori estimates for higher order finite elements for elliptic interface problems. *Applied Numerical Mathematics*, 60:19–37, 2009.

[60] Z. Li. A fast iterative algorithm for elliptic interface problems. *SIAM Journal on Numerical Analysis*, 35(1):230–254, 1998.

[61] Z. Li and K. Ito. Maximum Principle Preserving Schemes for Interface Problems. *SIAM J. Sci. Comput*, 23:339–361, 2001.

[62] Z. Li, T. Lin, and X. Wu. New Cartesian grid methods for interface problems using the finite element formulation. *Numerische Mathematik*, 96(1):61–98, 2003.

[63] K.-A. Lie. http://www.sintef.no/Projectweb/GeoScale/Results/MsMFEM/SPE10/. online. accessed 18th May 2011.

[64] T. Liu, B. Khoo, and C. Wang. The ghost fluid method for compressible gas-water simulation. *Journal of Computational Physics*, 204(1):193–221, 2005.

[65] T. Liu, B. Khoo, and K. Yeo. Ghost fluid method for strong shock impacting on material interface. *Journal of Computational Physics*, 190(2):651–681, 2003.

[66] X. Liu, R. Fedkiw, and M. Kang. A boundary condition capturing method for Poisson's equation on irregular domains. *Journal of Computational Physics*, 160(1):151–178, 2000.

[67] X. Liu and T. Sideris. Convergence of the ghost fluid method for elliptic equations with interfaces. *Mathematics of Computation*, 72(244):1731–1746, 2003.

[68] W. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge Univ Pr, 2000.

[69] N. Moes, J. Dolbow, and T. Belytschko. A finite element method for crack growth without remeshing. *International Journal for Numerical Methods in Engineering*, 46(1):131–150, 1999.

[70] S. Mousavi, E. Grinspun, and N. Sukumar. Harmonic enrichment functions: A unified treatment of multiple, intersecting and branched cracks in the extended finite element method. *International Journal for Numerical Methods in Engineering*.

[71] S. Mousavi, E. Grinspun, and N. Sukumar. Higher-order extended finite elements with harmonic enrichment functions for complex crack problems. 2010.

[72] J. Nolen, G. Papanicolaou, and O. Pironneau. A framework for adaptive multiscale methods for elliptic problems. *Multiscale Model. Simul*, 7(1):171–196, 2008.

[73] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, 1988.

[74] H. Owhadi and L. Zhang. Homogenization of the acoustic wave equation with a continuum of scales. *Arxiv preprint math.NA/0604380*, 2006.

[75] H. Owhadi and L. Zhang. Metric-based upscaling. *Communications on Pure and Applied Mathematics*, 60(5):675–723, 2007.

[76] C. Peskin. Numerical analysis of blood flow in the heart. *Journal of Computational Physics*, 25(3):220–252, 1977.

[77] L. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Mathematics of Computation*, 54(190):483–493, 1990.

[78] J. Sethian and A. Wiegmann. Structural boundary design via level set and immersed interface methods. *Journal of Computational Physics*, 163(2):489–528, 2000.

[79] T. Strouboulis, K. Copps, and I. Babuška. The generalized finite element method: an example of its implementation and illustration of its performance. *International Journal for Numerical Methods in Engineering*, 47(8):1401–1417, 2000.

[80] T. Strouboulis, L. Zhang, and I. Babuka. Generalized finite element method using mesh-based handbooks: application to problems in domains with many voids. *Computer Methods in Applied Mechanics and Engineering*, 192(28-30):3109–3161, 2003.

[81] T. Strouboulis, L. Zhang, and I. Babuška. p-version of the generalized FEM using mesh-based handbooks with applications to multiscale problems. *Int. J. Numer. Methods Eng*, 60:1639–1672, 2004.

[82] N. Sukumar, D. Chopp, N. Moës, and T. Belytschko. Modeling holes and inclusions by level sets in the extended finite-element method. *Computer Methods in Applied Mechanics and Engineering*, 190(46-47):6183–6200, 2001.

[83] S. Unverdi and G. Tryggvason. A front-tracking method for viscous, incompressible, multi-fluid flows. *Journal of Computational Physics*, 100(1):25–37, 1992.

[84] H. Wang and M. Anderson. *Introduction to groundwater modeling: finite difference and finite element methods*. Freeman, 1982.

[85] M. Wang, X. Wang, and D. Guo. A level set method for structural topology optimization. *Computer Methods in Applied Mechanics and Engineering*, 192(1-2):227–246, 2003.

[86] S. Wang and M. Wang. A moving superimposed finite element method for structural topology optimization. *International Journal for Numerical Methods in Engineering*, 65(11):1892–1922, 2006.

[87] P. Wei, M. Wang, and X. Xing. A study on X-FEM in continuum structural optimization using a level set model. *Computer-Aided Design*, 42(8):708–719, 2010.

[88] A. Wiegmann and K. Bube. The explicit-jump immersed interface method: finite difference methods for PDEs with piecewise smooth solutions. *SIAM Journal on Numerical Analysis*, pages 827–862, 2000.

[89] X. Wu, Y. Efendiev, and T. Hou. Analysis of upscaling absolute permeability. *Discrete and Continuous Dynamical Systems Series B*, 2(2):185–204, 2002.

[90] M. Yim and S. Simonson. Performance assessment models for low level radioactive waste disposal facilities: a review. *Progress in Nuclear Energy*, 36(1):1–38, 2000.

[91] L. Zhang. http://people.maths.ox.ac.uk/zhang/research.html. online. accessed 18th May 2011.

[92] Y. Zhou, S. Zhao, M. Feig, and G. Wei. High order matched interface and boundary method for elliptic equations with discontinuous coefficients and singular sources. *Journal of Computational Physics*, 213(1):1–30, 2006.

A

Elementary results on linear approximation

In this appendix we explore some of the linear approximation results required for Chapter 2. These results are essentially known but we recap and apply them to the work in this thesis for completeness. The idea is to show that shape regularity of a domain is preserved under a shape regular affine transformation. Note that throughtout this appendix we use the notation of Chapter 2, as such Ω is a domain in $\mathbb{R}^2$. Our first lemma bounds the maximum diameter of a domain in τ under the pullback F_τ^{-1}.

Lemma A.1. *For a triangular element σ and corresponding affine map F_σ (see Definition 2.30) we have that for a domain $\gamma \subset \Omega$*

$$H_{\hat{\gamma}} \lesssim \frac{H_\gamma}{\rho_\sigma} \tag{A.1}$$

where $\hat{\gamma} = \left\{\hat{x} \in \mathbb{R}^2 \mid F_\sigma(\hat{x}) \in \gamma\right\}$ is the pullback of γ.

Proof.

$$\begin{aligned} H_{\hat{\gamma}} = \max_{\hat{x}_1,\hat{x}_2\in\hat{\gamma}} |\hat{x}_1 - \hat{x}_2|_2 &= \max_{x_1,x_2\in\gamma} \left|F_\sigma^{-1}(x_1) - F_\sigma^{-1}(x_2)\right|_2 \\ &= \max_{x_1,x_2\in\gamma} \left|A_\sigma^{-1}x_1 - A_\sigma^{-1}x_2\right|_2 \\ &\lesssim \left|A_\sigma^{-1}\right|_2 \max_{x_1,x_2\in\gamma} |x_1 - x_2|_2 \\ &\lesssim \rho_\sigma^{-1} H_\gamma \end{aligned}$$

using Lemma 2.31. □

Our second lemma bounds the size of the largest inscribed ball under the action of the pullback.

Lemma A.2. *For a triangular element $\sigma \in \mathcal{T}_H(\Omega)$ and a domain γ we have*

$$\rho_\gamma \lesssim H_\sigma \rho_{\hat{\gamma}} \tag{A.2}$$

where $\hat{\gamma}$ is the pullback of γ under F_σ^{-1}.

Proof. For this proof we are essentially interested in the action of an affine map on a ball. To this end we need to show that a circle in $\mathbb{R}^2$ maps to an ellipse under an affine transform. Consider the canonical formula for a circle,

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = r^2.$$

Under the affine transform F_σ this becomes

$$\begin{bmatrix} \hat{x} & \hat{y} \end{bmatrix} A_\sigma^T A_\sigma \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} + 2b_\sigma^T A_\sigma \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} + b_\sigma^T b_\sigma = r^2. \tag{A.3}$$

for $\hat{x}, \hat{y} \in \mathbb{R}^2$ which is in the form of a conic section whos general formula is

$$\begin{bmatrix} \hat{x} & \hat{y} \end{bmatrix} \begin{bmatrix} w_1 & \frac{w_2}{2} \\ \frac{w_2}{2} & w_3 \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} + \begin{bmatrix} w_4 & w_5 \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} + w_6 = 0.$$

Now this is an ellipse if the descriminant $w_2^2 - 4w_1w_3 < 0$. However from (A.3)

$$w_2^2 - 4w_1w_3 = -4\text{det}\left(A_\sigma^T A_\sigma\right).$$

From Definition 2.30 we have that

$$\det\left(A_\sigma^T A_\sigma\right) = |(y_2 - y_1) \times (y_3 - y_1)|_2^2 = 4\,|\sigma|^2 > 0\ .$$

This means the descriminant $w_2^2 - 4w_1w_3 = -4\text{det}\left(A_\sigma^T A_\sigma\right) < 0$ and therefore a circle is mapped to an ellipse under an affine transform.

Now for the proof of the lemma. Let ρ_γ be the diameter of the largest inscribed ball in γ, denote the boundary of this ball by ∂B_γ. Under the mapping F_σ^{-1}, ∂B_γ will be mapped back to an inscribed ellipse $\partial E_{\hat{\gamma}} \subset \hat{\gamma}$, where $\hat{\gamma}$ is the pullback of γ under F_σ^{-1}.

Let $\partial E_{\hat{\gamma}}$ have a semimajor and semiminor axis, shown in Figure A-2 as the vectors between $\hat{a}_1$ and $\hat{a}_2$ and then $\hat{b}_1$ and $\hat{b}_2$ respectively.

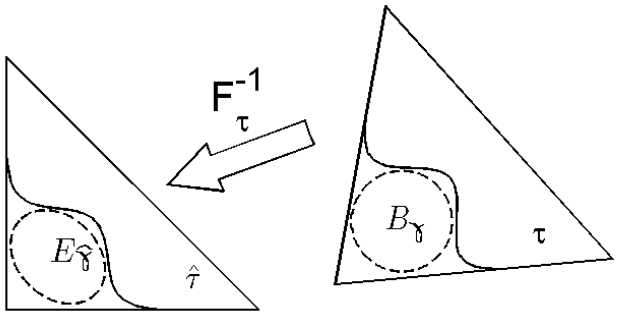


Figure A-1: *An example of how a circle is mapped to an ellipse under an affine map.*

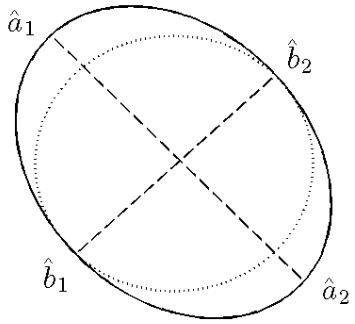


Figure A-2: *How $\partial B_{\hat{\gamma}}$ is defined inside the ellipse created from the pullback of ∂B_{γ}.*

Take the points $\hat{b}_1$ and $\hat{b}_2$ as defining the diameter of an inscribed circle in $\hat{\gamma}$, not necessarily the largest inscribed circle. This implies

$$\left\|\hat{b}_1 - \hat{b}_2\right\|_2 \leq \rho_{\hat{\gamma}} \ . \tag{A.4}$$

We now combine all this together to get the final result. Using $\hat{b}_1$, $\hat{b}_2$ and Lemma 2.31 and (A.4) we get

$$\begin{aligned}\rho_{\gamma} = \left\|F_{\sigma}(\hat{b}_1) - F_{\sigma}(\hat{b}_2)\right\|_2 &= \left\|A_{\sigma}\hat{b}_1 - A_{\sigma}\hat{b}_2\right\|_2 \\ &\lesssim |A_{\sigma}|_2 \left\|\hat{b}_1 - \hat{b}_2\right\|_2 \\ &\lesssim H_{\sigma}\rho_{\hat{\gamma}}\end{aligned}$$

as required. □

As a consequence of these two previous lemmas we obtain a corollary that shows shape regularity is preserved for a shape regular affine transform.

Corollary A.3. *Suppose σ is a shape regular triangular element (see Assumption 2.15) with corresponding affine map F_σ. For a domain $\gamma \subset \Omega$ we have that*

$$\frac{H_{\hat{\gamma}}}{\rho_{\hat{\gamma}}} \lesssim \frac{H_\gamma}{\rho_\gamma} \ . \tag{A.5}$$

Proof. From Lemma A.2 we know that $\rho_\gamma \lesssim H_\tau \rho_{\hat{\gamma}}$ so if we divide both sides by $\rho_\gamma \rho_{\hat{\gamma}}$ we get

$$\frac{1}{\rho_{\hat{\gamma}}} \lesssim \frac{H_\tau}{\rho_\gamma} \ .$$

Then from Lemma A.1 we have that

$$H_{\hat{\gamma}} \lesssim \frac{H_\gamma}{\rho_\tau} \ .$$

Combining these last two equations we get

$$\frac{H_{\hat{\gamma}}}{\rho_{\hat{\gamma}}} \lesssim \frac{H_\gamma}{\rho_\gamma} \frac{H_\tau}{\rho_\tau} \lesssim \frac{H_\gamma}{\rho_\gamma} \ .$$

□